THE BELL SYSTEM Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXIX SEPTEMBER 1960 NUMBER 5

Binocular Depth Perception of Computer-Generated Patterns B. JULESZ 1125

Models for Approximating Basilar Membrane Displacement J. L. FLANAGAN 1163

Design and Performance of Ultraprecise 2.5-mc Quartz Crystal Units A. W. WARNER 1193

Some Further Theory of Group Codes D. SLEPIAN 1219

Capacity of a Burst-Noise Channel E. N. GILBERT 1253

Automata and Finite Automata C. Y. LEE 1267

Transition Probabilities for Telephone Traffic V. E. BENEŠ 1297

An Alternative Approach to the Realization of Network Transfer Functions: The *N*-Path Filter L. E. FRANKS AND I. W. SANDBERG 1321

Magnetic Latching Crossbar Switches: A New Development in Magnetic Properties of Tool Steel F. A. ZUPA 1351

Recent Bell System Monographs 1375

Contributors to This Issue 1379

COPYRIGHT 1960 AMERICAN TELEPHONE AND TELEGRAPH COMPANY

THE BELL SYSTEM TECHNICAL JOURNAL

ADVISORY BOARD

H. I. ROMNES, *President, Western Electric Company*

J. B. FISK, *President, Bell Telephone Laboratories*

E. J. McNEELY, *Executive Vice President, American Telephone and Telegraph Company*

EDITORIAL COMMITTEE

A. C. DICKIESON, *Chairman*
S. E. BRILLHART
A. J. BUSCH
L. R. COOK
R. L. DIETZOLD
J. H. FELKER
K. E. GOULD
E. I. GREEN
G. GRISWOLD, JR.
J. R. PIERCE
M. SPARKS
W. O. TURNER

EDITORIAL STAFF

W. D. BULLOCH, *Editor*
R. M. FOSTER, JR., *Assistant Editor*
C. POLOGE, *Production Editor*
J. T. MYSAK, *Technical Illustrations*
T. N. POPE, *Circulation Manager*

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. F. R. Kappel, President; S. Whitney Landon, Secretary; L. Chester May, Treasurer. Subscriptions are accepted at $5.00 per year. Single copies $1.25 each. Foreign postage is $1.08 per year or 18 cents per copy. Printed in U.S.A.

THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XXXIX SEPTEMBER 1960 NUMBER 5

Copyright 1960, American Telephone and Telegraph Company

Binocular Depth Perception of Computer-Generated Patterns

By BELA JULESZ

(Manuscript received March 31, 1960)

The perception of depth involves monocular and binocular depth cues. The latter seem simpler and more suitable for investigation. Particularly important is the problem of finding binocular parallax, which involves matching patterns of the left and right visual fields. Stereo pictures of familiar objects or line drawings preclude the separation of interacting cues, and thus this pattern-matching process is difficult to investigate. More insight into the process can be gained by using unfamiliar picture material devoid of all cues except binocular parallax. To this end, artificial stereo picture pairs were generated on a digital computer. When viewed monocularly, they appear completely random, but if viewed binocularly, certain correlated point domains are seen in depth. By introducing distortions in this material and testing for perception of depth, it is possible to show that pattern-matching of corresponding points of the left and right visual fields can be achieved by first combining the two fields and then searching for patterns in the fused field. By this technique, some interesting properties of this fused binocular field are revealed, and a simple analog model is derived. The interaction between the monocular and binocular fields is also described. A number of stereo images that demonstrate these and other findings are presented.

I. INTRODUCTION

The question of how the two-dimensional projections of the visual world that are supplied to the left and right eyes are matched and combined to reveal the impression of depth is an extremely interesting one. Because of an incorrect analogy derived from measuring distances with a range finder, it is commonly thought that this problem is rather trivial. Admittedly, it is fairly simple to determine binocular parallax by aligning selected portions of an object in the left and right fields of a range finder and computing depth by trigonometrical calculations. The intriguing part of this problem is to explain the remarkable ability of humans to establish *correspondence* between complicated patterns in the two monocular fields. This pattern-matching process is the one being investigated here.

It seems quite clear that patterns perceived in depth afford a promising means for exploring pattern-matching. However, it is well known that the perception of depth under familiar conditions is mediated by many complex cues, both binocular and monocular, which are not easily kept under the control of the experimenter. Thus, many previous explorations have used stereo pictures of familiar objects or line drawings, precluding the separation of interacting cues. The investigation reported here utilized patterns devoid of all cues except binocular parallax, by using artificially created stereo images with known topological properties. Such visual displays ordinarily never occur in real-life situations, and a digital computer (with a video transducer at its output) was programmed to generate them. When these unfamiliar pictures are viewed stereoscopically, peculiar and often unexpected depth effects can be seen. In addition, the perception time of depth under such circumstances is sometimes in the order of minutes (instead of the few milliseconds required for familiar stereo images). This slowing down of the visual process facilitated the present investigation without having much effect on the stability of depth impression after depth was finally perceived.

This paper reports a study of binocular depth perception based upon such presentations. In Section II the problem is posed explicitly and a summary of the results is given. The intent is to provide the essence of this investigation without going into details. The remaining sections are arranged along the sequence of ideas presented in Section II, with the intention of being more specific and of supplying more data. In the last section the new technique of this investigation is evaluated with some possible future applications.

A pair of Fresnel lenses has been enclosed on page 1161 of this issue of the Bell System Technical Journal. They may be used for viewing the stereoscopic illustrations in this paper. Directions for their use may be found in the Appendix.

II. PROBLEM POSING AND SUMMARY OF RESULTS

Human beings exhibit great ability in utilizing binocular parallax to establish the relative depth of objects in the visual field. This process involves finding horizontal shifts between corresponding point domains in the left and right visual fields. The observer seems able to establish this correspondence almost without effort or deliberation, even when the fields differ in brightness and shape (due to reflections and perspective) and in picture material (due to hidden objects seen by only one eye). Thus, depth perception might be likened to the solution of a complicated pattern-recognition problem.

This paper attacks the problem of depth perception as a pattern-recognition problem and poses the following question: In determining binocular parallax do we first recognize monocular patterns in the left and right fields and then fuse them (monocular pattern recognition), or do we first combine the two fields in some manner and then perform all further processings on the fused field [e.g., search for certain patterns (binocular pattern recognition)], or do we utilize a combination of both processes? This question is appropriate both for macropatterns (higher organization of points into objects) and for micropatterns (a few adjacent points). Figs. 1, 2 and 3 attempt merely to illustrate these three possibilities and do not necessarily have relevance to physiological systems.

Artificial stereo images were created by an IBM 704 digital computer. Right and left images were generated, each consisting of 10,000 brightness points, which were assigned one of 16 quantized brightness values at random. In a peripheral "surround" region, the images were identical; in a square-shaped central region, the right-hand image differed from the left by a uniform horizontal displacement. When viewed monocularly, the images appear completely random. But when viewed stereoscopically, this image pair gives the impression of a square markedly in front of (or behind) the surround. By fusing the photographs in Fig. 4 (using two lenses as prisms with a diameter of 2 inches or more and 10 to 18

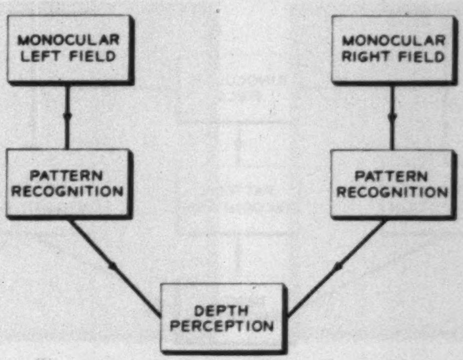


Fig. 1 — Depth perception by monocular pattern recognition.

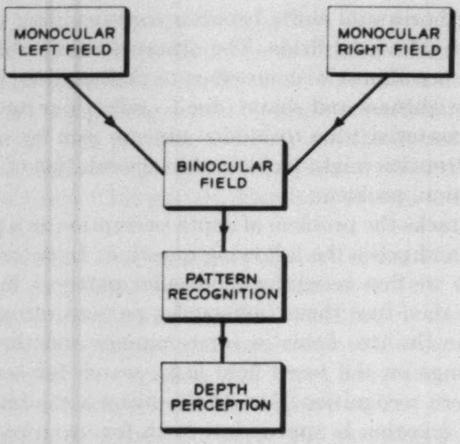

MONOCULAR LEFT FIELD
MONOCULAR RIGHT FIELD
BINOCULAR FIELD
PATTERN RECOGNITION
DEPTH PERCEPTION

Fig. 2 — Depth perception by binocular pattern recognition.

inches focal length, such as those supplied with this issue) this depth effect can be demonstrated.

Of course, depth perception under these conditions takes longer to establish because of the absence of monocular cues. Still, once depth is perceived, it is quite stable. This experiment shows quite clearly that it is possible to perceive depth without monocular macropatterns. However, if binocular pattern recognition is the principal depth mechanism, the same statement should be true for micropatterns.

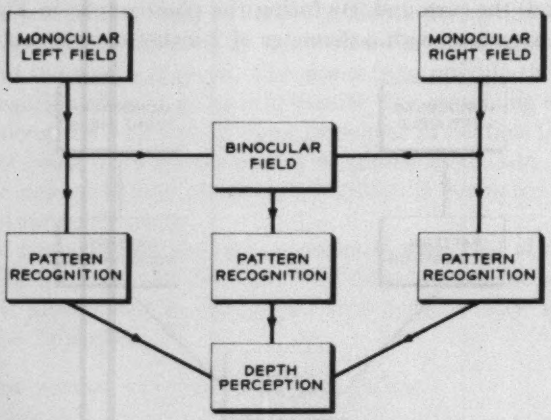

MONOCULAR LEFT FIELD
MONOCULAR RIGHT FIELD
BINOCULAR FIELD
PATTERN RECOGNITION
PATTERN RECOGNITION
PATTERN RECOGNITION
DEPTH PERCEPTION

Fig. 3 — Depth perception by monocular and binocular pattern recognition.

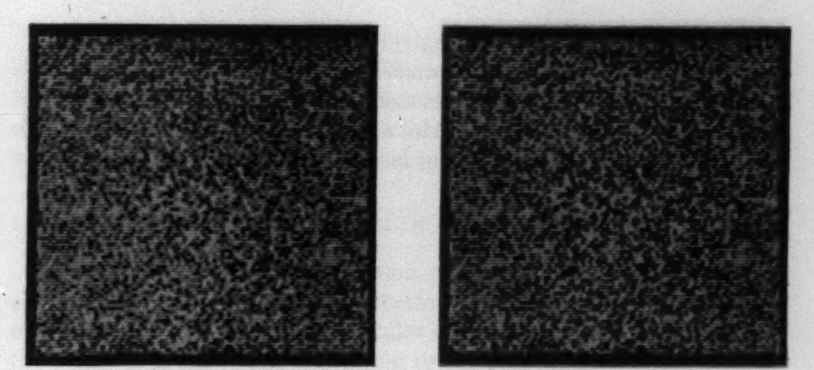


Fig. 4 — Stereo pair with center square above the background.

To study this matter, micropatterns in the stereo pair were drastically altered by blacking out a regular pattern of points in the left field and making the corresponding points white in the right field. Fig. 5 shows the result of this process, where the perturbation grid consists of every even point of every even line. The microstructure of the left and right images is highly different, and yet the center square stands out clearly from the surround.

In spite of the difference in microstructure of the left and right fields, this experiment may not be decisive. It could be argued that the regular perturbation grid is recognized monocularly in its random surround and disregarded, and that the remaining, unaltered points in the two fields possess the same microstructure. It was found, however, that the difficulty of monocularly recognizing the perturbation grid could be increased

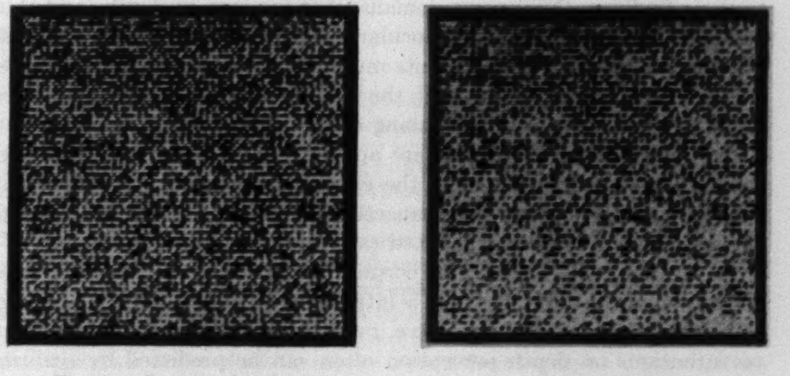


Fig. 5 — Stereo pair with superimposed unmixed perturbation grid.

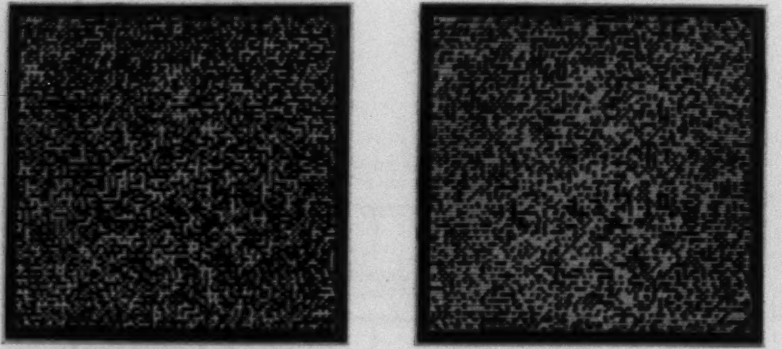


Fig. 6 — Stereo pair as in Fig. 5, but quantized into two levels.

greatly without increasing the difficulty of perceiving depth. For instance, when the random fields are quantized into two levels (black and white), the perturbation grid composed of black (or white) points seems more difficult to find in this surround than in one composed of 16 brightness levels (with many medium grays). The depth effect in Fig. 6 (two-level quantization) can be obtained with the same ease as it can in Fig. 5 (16-level quantization). This makes the assumption of monocularly recognizing the grid very improbable. Together with other evidence (to be discussed in Section VII), it therefore suggests strongly that the two fields are combined first and that the processing is done on the fused field.

Other experiments making use of similar techniques are described. The results shed light on pattern recognition as it is involved in binocular vision. The problem of detecting certain regions in the fused binocular field in order to find depth was particularly investigated. According to these findings, those point domains that are seen in depth (and thus have to be detected in the binocular field) need not possess a *Gestalt*, but the connectivity of the points must be preserved. In the above-described regular perturbation grid, the unaltered points are still connected along one-dimensional arrays (along every other line and column). But if meshlike perturbation grids are applied (which leave the same per cent of points unaltered as in the experiments that will be shown in Figs. 20 and 21, but limit the connectivity of points to small subregions), the depth effect is greatly reduced (as will be seen in Figs. 26 and 27).

As an interesting analogy to certain properties of the binocular field the notion of the difference field is introduced (see Section IX). Although this model is probably very naïve, nevertheless the influence of various perturbations on depth perception often can be predicted by realizing some trivial properties of the corresponding difference field.

The concluding experiments investigate the role of monocular macropatterns in depth perception. It is shown that their presence greatly enhances the depth effect; thus, monocular and binocular pattern recognition can occur simultaneously as a mixed process. This statement seems to be the final answer to the original problem.

III. BRIEF EVALUATION OF MONOCULAR AND BINOCULAR DEPTH CUES

Depth perception is an interaction of extremely complicated mental processes. These processes utilize certain depth cues which usually are divided into binocular and monocular depth cues. In Table I, a list of these cues is given (without aspiring to completeness).[1]

Most of the monocular depth cues require a tremendous memory capacity; for instance, familiarity with perceived objects implies a catalog of no mean extent.

Binocular depth cues seem simpler and more akin to data processing. Binocular convergence and accommodation are very weak depth cues (as tachistoscopic experiments[2] have shown), and they can be ignored in favor of binocular parallax, which is apparently the principal binocular cue. The invention of the stereoscope[3] strikingly demonstrated man's ability to utilize binocular parallax in order to perceive depth — that is, to determine correspondence between points in the left and right visual fields and measure the horizontal displacements between them. The importance of monocular depth cues in supplementing binocular depth cues is great, as can be demonstrated by the reversed depth effect. It is well

TABLE I

Binocular Depth Cues
Binocular parallax
Convergence of eyes
Correlative accommodation (focusing)
Monocular Depth Cues
Linear perspective (such as converging railroad tracks)
Apparent size of objects of known size (which decreases with distance of observer)
Monocular parallax (change of appearance with change of observer's position)
Shadow patterns (the light-and-shade relations yielding relief)
Interposition (the superposing of near objects on far objects)
Changes due to atmospheric conditions (such as haze, blurring of outlines)
Accommodation (focusing on an object with one eye)
Retinal gradient of texture (decreasing size of texture elements with distance)
Retinal gradient of size of similar objects (rate of decrease of size of houses, fence posts, telegraph poles, etc.)

known that, by interchanging the left and right picture pair in a stereoscope, unfamiliar objects reverse their depth coordinates (far points become near, convex surfaces become concave, etc.). For a familiar object (e.g., a human face) the reversal of depth relationships usually does not take place; that is, the monocular depth cues counteract the binocular ones.

To cancel the effect of these involved monocular depth cues and concentrate on the binocular parallax, most work with stereoscopes uses line drawings for visual stimulus. These drawings comprise simple dots, lines, circles, etc., with different parallax shifts in the right and left fields, and are practically free of monocular depth cues. A vivid depth effect can still be obtained.

The above-mentioned tachistoscopic experiment deserves some additional explanation. A stereo pair consisting of simple line drawings (with parallax shifts in nasal or temporal directions) was flashed for a brief period (in the order of a few milliseconds). Viewing it stereoscopically, subjects could tell almost without any error which of the drawings were in front of or behind a reference plane. This experiment tells nothing about the time required to perceive depth because of the long-persistent afterimages, but it gives some insight into depth processes. First of all, during the short exposure period no convergence or any other motion of the eyes can take place. This fact excludes convergence and accommodation as important depth cues. Second, it demonstrates that during fusion the left and right fields must be labeled, because otherwise the perception of near and far would be confused.

The following investigations are based on the possibility of separating the monocular and binocular depth cues, and concentrate on the problem of how binocular parallax can give the impression of depth.

IV. MACROPATTERN AND MICROPATTERN RECOGNITION; MONOCULAR AND BINOCULAR PATTERN RECOGNITION

It seems clear that a basic aspect of depth perception is recognition of binocular parallax, which consists of a parallax shift between corresponding points in the left and right visual fields. The shift is parallel to the base line (of the eyes); thus, the corresponding points in the left and right fields must lie on the same horizontal line. Now, to determine the exact amount of parallax shift, it is necessary to find the corresponding points in the left and right visual fields. Because the base distance (between the two eyes) and the focal length of the eyes (looking at the stereo pictures at a given distance) are known, there is a simple trigonometric relationship between the parallax shift and the actual depth.

Thus, determining the parallax for every point is analogous to the reconstruction of three-dimensional space. So we come to the kernel of the problem: How can we fuse points in the left and right fields and establish correspondence between them in a stereo sense, when the two fields may differ quite drastically from each other?

The left and right fields of a stereo pair can differ: (a) in brightness (due to different reflections); (b) in perspective (expansion, rotation, shift, etc., of point domains); and (c) by hidden parts (seen only by one eye). Obviously, one is able somehow to find the points in the two fields that belong to considerably different patterns. How is this equivalence established? Do we recognize a face, a square, a few adjacent points, etc., in the left and right visual fields separately and then pick up the corresponding points, or do we first fuse the two fields and perform certain pattern-recognition tasks on this fused field?

To make these questions more precise we introduce the following terminology: Pattern recognition can be divided into two classes. First, *micropattern recognition* concerns simple pattern organizations that take into account some geometrical, topological characteristics in a point's immediate neighborhood. Second, *macropattern recognition* is a higher-order organization of several points. Points grouped together and recognized as a face, square, number, etc., are examples of what is meant by this conception.

The first half of another useful dichotomy is *monocular pattern recognition*, which is performed on the visual field seen by one eye. *Binocular pattern recognition* is performed on the fused field, which is a combination of the left and right monocular fields. It belongs to a special class of processings that incorporate characteristics that intuitively are also important in ordinary (monocular) pattern recognition. Nevertheless, binocular pattern recognition need not necessarily be identical or even similar to monocular pattern recognition.

With these distinctions in mind, we may ask: Is the basic mechanism of binocular fusion a monocular pattern recognition (Fig. 1), or a binocular pattern recognition (Fig. 2), or a combination of both (Fig. 3)? These possibilities multiply when we further differentiate between micropattern and macropattern recognition in each case.

V. DEPTH PERCEPTION WITHOUT MONOCULAR MACROPATTERN RECOGNITION

In aerial reconnaissance it is known that objects camouflaged by a complex background are very difficult to detect monocularly but jump out if viewed stereoscopically. Though the macropattern (hidden object) is *difficult* to see monocularly, it *can* be seen. Therefore, this evidence is not sufficient to prove that depth can be perceived without monocular macropattern recognition.

To investigate this problem, a special visual presentation was created by means of the IBM 704 digital computer and a television transducer developed in the Visual and Acoustics Research Department of Bell Telephone Laboratories.[4,5,6] A pseudo random number routine was programmed to generate random numbers in sequence according to a uniform probability distribution. These numbers were quantized in 16 levels, which were written on tape and then translated by means of a digital-to-analog converter and a special television scanner into 16 brightness levels between black and white. (The quality of present scanning techniques and of photographic processes limits the resolution in brightness, and the final pictures have actually less than 16 identifiable levels.) The television scanner used has the format of a two-dimensional rectangular matrix of 99 rows, each consisting of 105 picture elements. Thus, a picture consists of 105 × 99 = 10,395 points, whose brightness assume randomly any of the 16 values between the maximum black and white.

A left- and a right-hand stereo image are created by the above-mentioned technique in the following way:

In a peripheral "surround" region, the right- and left-hand images are identical (i.e., the same random brightness points are copied in the two pictures in the same locations); however, in a square-shaped central region, the right image differs from the left by a uniform horizontal displacement. Fig. 7 illustrates this procedure on a small matrix of 6 × 6 elements. The background points are indicated with small letters having a range of eight letters (brightness values) taken at random. The shifted square in the center has 2 × 2 elements (indicated by capital

a	b	a	c	d	f
g	e	h	d	c	b
e	f	A	G	a	g
e	a	D	B	e	c
f	c	d	e	f	e
d	g	c	h	b	a

a	b	a	c	d	f
g	e	h	d	c	b
e	A	G	c	a	g
e	D	B	d	e	c
f	c	d	e	f	e
d	g	c	h	b	a

Fig. 7 — Illustration of method by which stereo random pictures are generated.

letters), and the parallax shift in the right field is one picture element to the left.

The distinction between small and capital letters is only for illustration; they possess the same range and distribution, and therefore no macropattern can be seen on any of these random images viewed separately. Those points which are seen only by one eye [e.g., the right side of the square on the right image (*c*, *d*)] are generated by the same random number routine.

Fig. 4 showed a stereo pair of 99 × 105 picture elements, the hidden central square having 40 × 40 elements, and the parallax shift (Δ) being four picture elements. Both of these pictures, viewed separately, give an entirely random impression, and only an experiment can determine whether when fused stereoscopically the center square will be seen in depth in front of (or behind) the surround.

The images presented can be fused easily by using two simple lenses (of more than two-inch diameter and 10- to 18-inch focal length) as prisms. After fusion, there is a vivid depth effect. The square is in front of the background plane, and the depth impression is very stable. It is interesting that the depth effect does not appear at once, but appears only after a fairly long time in comparison to that in familiar stereo pictures. A curious learning process can be experienced; that is, the time required to get the depth effect diminishes after repetitive trials. The problem of what is really learned here is an interesting question in itself and deserves further investigation.

The fuzziness of the edges of the square is mainly due to the fact that, by chance, some of the brightness points along the edges of the square can belong to both the square and the background, and there is a tendency to interpret them ambiguously. The probability that two or more adjacent points should become ambiguous is very low, so the fuzziness of the edges is about ±1 picture element in width. (These rough edges reveal that no "*Gestalt* organization" takes place in binocular fusion though the square has a "good *Gestalt*.")

Fig. 8 demonstrates another stereo pair generated in the above way by the computer, but now there are three planes: the background plane, a central rectangle 60 × 40 in size and with a parallax shift of $\Delta_1 = 4$, and a third rectangle 20 × 40 in size and with a parallax shift of $\Delta_2 = 8$. It takes some time to get the bigger rectangle in front of the background, but it usually takes even more time to get the smaller rectangle in front of the bigger one. After the three different planes of depth are perceived they remain very stable. The same is true for the reversed depth effect. If the left and right images are interchanged (thus the parallax shifts

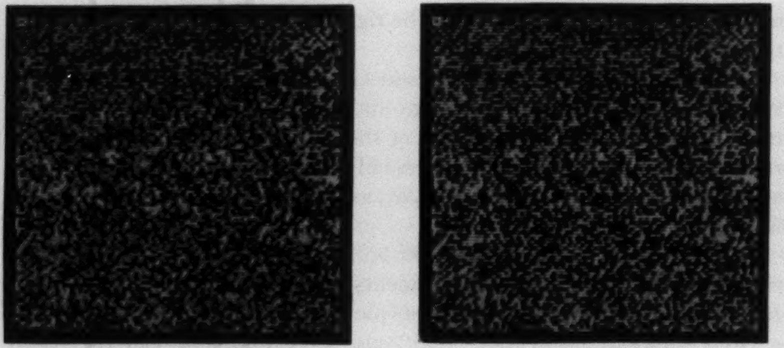


Fig. 8 — Stereo pair with two different planes of depth above background.

are not in the nasal direction but in the temporal one), the three planes reverse their depth relation to the observer. If highlights are eliminated and the surface-like appearance of the pictures is reduced, no monocular depth cues remain, and the reversed depth effect can be obtained with the same ease as in the regular case (Fig. 9).

Apparently, the greater difficulty in seeing the smaller rectangle at its "proper" depth arises, not because of its greater parallax shift, but merely because of its smaller size. By using the same parallax shifts as in Fig. 8 but increasing the size of the closest rectangle and decreasing the intermediate one, it can be demonstrated that the closest rectangle emerges first from the background followed by two smaller ones behind on the sides (see Fig. 10).

These experiments show that it is possible to perceive depth without

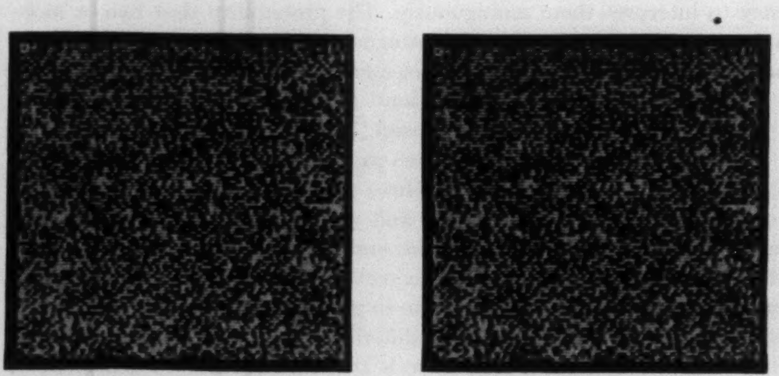


Fig. 9 — Stereo pair with two different planes of depth behind foreground.

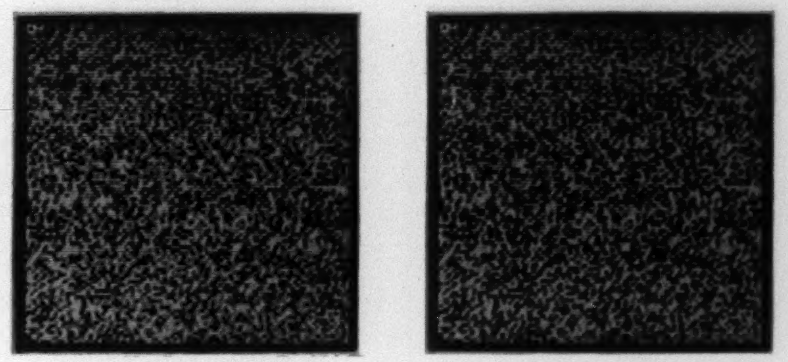


Fig. 10 — Stereo pair with two different planes of depth above background.

monocular macropattern recognition. We must now investigate this same matter for micropattern recognition, if the flow chart for depth perception is to be established. In Sections VI and VII this problem is investigated.

VI. EFFECTS OF INTRODUCED PERTURBATIONS ON THE DEPTH PERCEPTION OF STEREO RANDOM FIELDS

If we compare ordinary stereo photographs of real-life objects, the left and right pictures can differ substantially without being difficult to fuse.

In the present investigation we concentrate only on local perturbations, such as differences in brightness, and ignore the problem of differences in perspective (expansions, rotations, etc.), which belongs to the class of perturbations extending over the pictures according to complicated laws.

The perturbations were introduced in only one of the two pictures, leaving the other unchanged. The perturbations naturally have an effect on the general appearance of the fused image and on the stability of depth perception, but these are not really the effects we are interested in. Our basic question was to find out whether or not, after a given type and amount of perturbation, depth could still be perceived. In other words, to what extent can the brain solve the problem of pattern-matching after distortions are introduced?

In the following investigations some limitations are imposed on the input material. The random stereo images contain only point domains with a uniform parallax shift. The value of the parallax shift and the

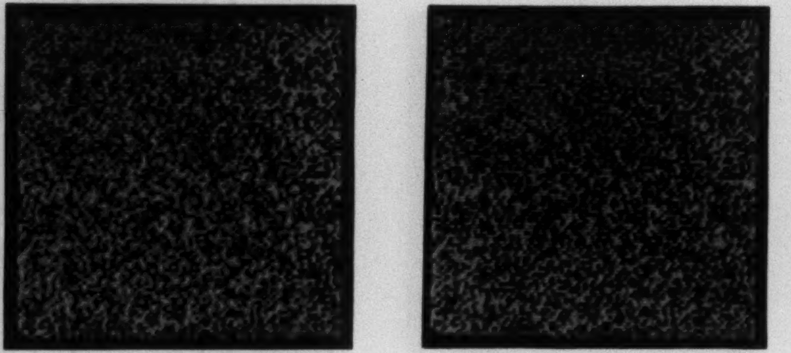


Fig. 11 — Stereo pair with gaussian noise perturbation (14-db signal to noise).

size of the center square is kept constant. The stereo pair before perturbation is like Fig. 4, i.e., two 40 × 40 squares with $\Delta = 4$.

The first type of perturbation introduced was the addition of gaussian noise on one of the stereo images. In Fig. 11 gaussian noise is added to the left picture. The signal-to-noise ratio (peak-to-peak signal to average noise) is 14 db. Nevertheless, the square is clearly visible in depth though several ambiguous points on the background and the square give rise to a lacy appearance. Even with a perturbation of 6 db signal to noise, the depth effect can be obtained, although the image is markedly deteriorated. Some additional findings will be discussed in Section IX.

Another type of noise is introduced by quantizing one of the stereo pairs in fewer levels than the other image. In Fig. 12 the left picture is

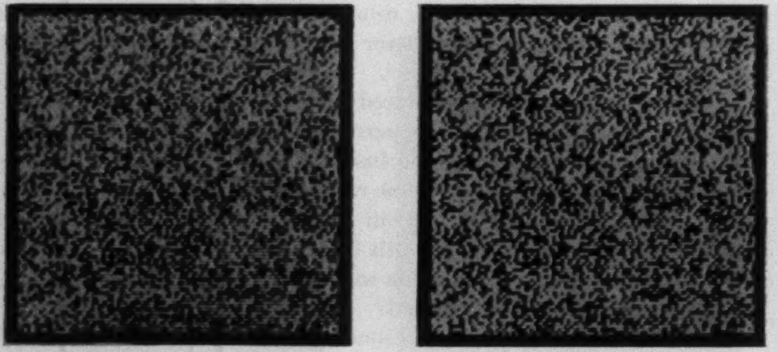


Fig. 12 — Stereo pair with quantizing noise perturbation.

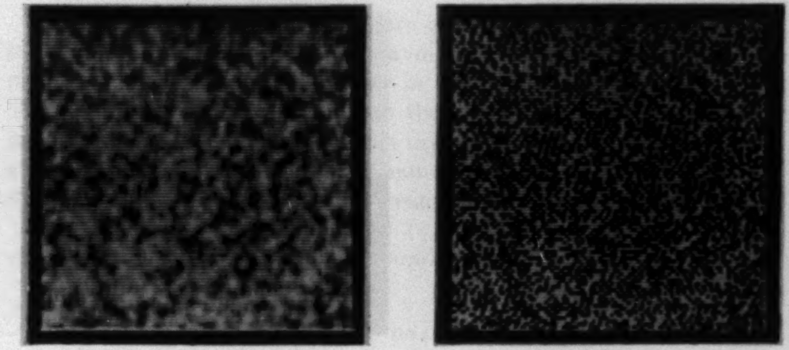


Fig. 13 — Stereo pair with blurred left picture.

quantized only into two levels (black and white). A decision level in the middle gray was chosen, and whenever a brightness point was greater than this it was represented as white, otherwise as black. The right picture is not altered and has 16 brightness levels (actually, on the photograph reproduced here, it has less, but there are more than four). This perturbation, in effect, yields to a special type of noise, sometimes called quantizing noise, and by fusing the stereo pair of Fig. 12 it becomes apparent that even this disturbance does not cancel the depth effect.

The next experiment uses a random stereo pair similar to Fig. 4 (but both the left and right images are quantized into two levels), and the left image is blurred (see Fig. 13). The blur is introduced in the computer by taking each point of the original image and adding to it its surrounding points with equal weights. The blurred u_i^* brightness points of Fig. 13 were obtained according to the following operation:

$$u_i^* = \tfrac{1}{9} \sum_{j=0}^{8} u_{ij} ,$$

using the notations in Fig. 14.

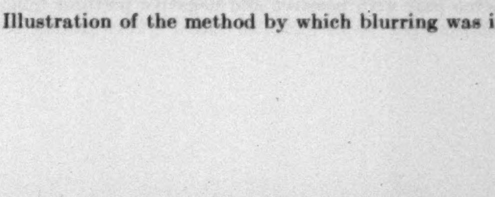


Fig. 14 — Illustration of the method by which blurring was introduced.

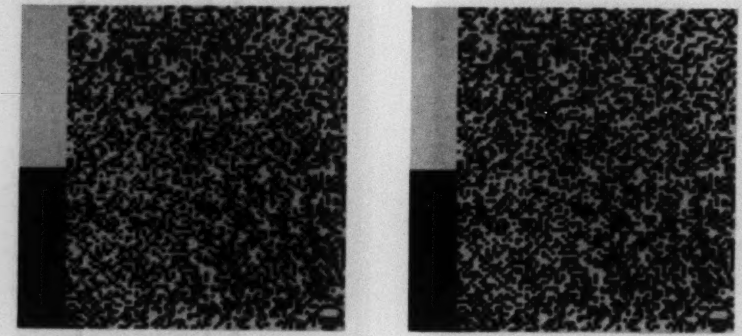


Fig. 15 — Stereo pair with positive pictures quantized into two levels.

This amount of blurring reduces the information content of the left image considerably, but it is still enough to carry the depth information. What is more, the eye is able somehow to see the whole as a sharp picture.

The following experiment is instructive in itself, and will be referred to in the next section. Fig. 15 shows a stereo pair (as does Fig. 4), but both left and right pictures are quantized into two levels. Depth can be easily perceived. Now in Fig. 16 the left picture is identical to the left picture in Fig. 15, but the right picture is the negative of the right picture in Fig. 15. Thus, all points are complemented. Experimenting with Fig. 16, we can conclude that it is not possible to fuse a positive and a negative picture. In addition, strong binocular rivalry can be experienced.

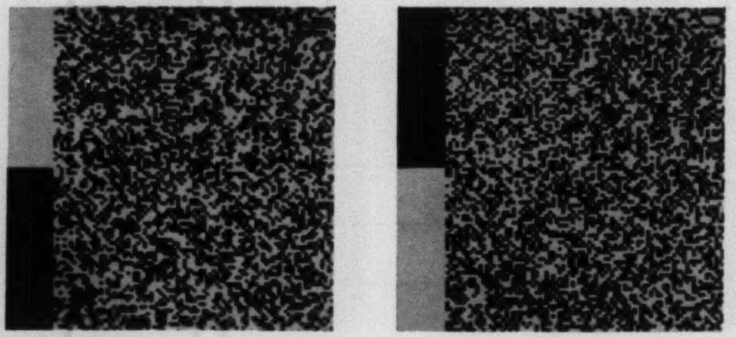


Fig. 16 — Stereo pair with positive and negative pictures quantized into two levels.

In these presentations, special care was taken to ensure uniformness of the black and white values. (To avoid filter ringing, we used a "sample and hold" circuit without filter in the digital-to-analog converter). The bars on the left side illustrate the effect of fusion and rivalry of more extended uniform areas. This experiment shows that one of the greatest perturbations we can introduce is to use maximum black or white points in one field and their complements in the other.

VII. DEPTH PERCEPTION WITHOUT MONOCULAR MICROPATTERN RECOGNITION

The perturbations introduced in the previous section were not drastic, and so the corresponding micropatterns in the left and right images still had some resemblance to each other. Nevertheless, it is apparent that fusion is not the result of a simple point-to-point correspondence between the stereo images. At least, certain coding operations that enhance the resemblances between corresponding micropatterns are required before fusion.

In the next experiments, the resemblance between the left and right micropatterns is drastically reduced; despite this fact, depth can be perceived in several situations.

In all the experiments that follow, the original stereo image is identical to the one in Fig. 4, with either 16 or two brightness levels and $\Delta = 4$. Then, a regular grid is superimposed on the left and right random fields, as shown in Fig. 17.

Every second point in every second line (shaded squares) is changed to maximum black in the left field and to maximum white in the right field. As shown, 25 per cent of the points are so treated, with the result that these points cannot be fused. The rest is unaltered. This arrangement of the perturbation grid removes similarities between the micropatterns of the stereo pairs in the following sense: There are not any corresponding points in the left and right images which have an identical neighborhood. At least one point is changed to its complement in any micropattern 2×2 or greater in size. Fig. 5 shows such a stereo pair of 16 brightness levels having a black grid in the left field and a corresponding white grid in the right field.

The grid cannot be seen monocularly, since it is embodied in the random field. When Fig. 5 is viewed binocularly, however, the square jumps out and is quite stable.

This experiment is still not decisive. One might argue that the resemblance between corresponding micropatterns is not completely removed because, along every other horizontal or vertical line (these are

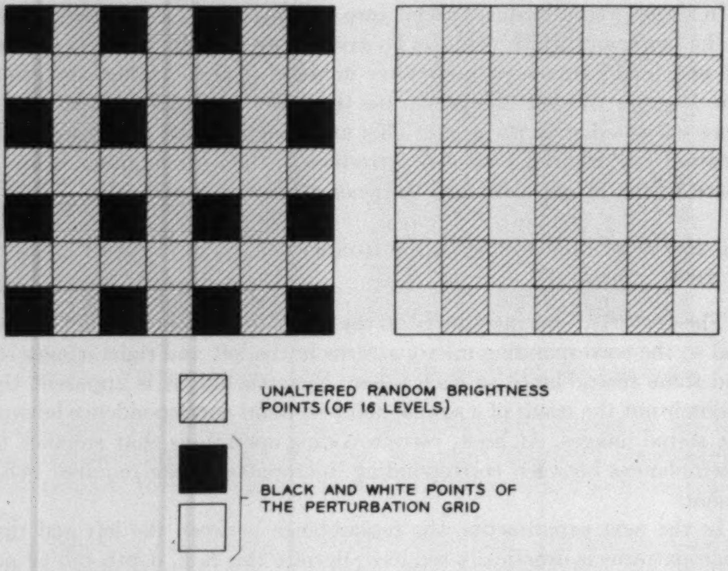

UNALTERED RANDOM BRIGHTNESS POINTS (OF 16 LEVELS)
BLACK AND WHITE POINTS OF THE PERTURBATION GRID

Fig. 17 — Illustration of the method by which the unmixed regular perturbation grid was generated.

unaltered), the micropatterns in the two fields are identical. A searching operation might exist that finds in the left and right monocular fields such identical one-dimensional arrays. To investigate this objection the following experiment was performed:

The same regular perturbation grid of Fig. 17 was used, but with a modification. Instead of uniformly blackening out all of the grid points in the left field, these points were made black or white at random. Then, the corresponding points in the right-hand field were assigned the complementary values (see Fig. 18).

Fig. 19 shows a 16-level random stereo field with this kind of mixed regular grid. Under these conditions depth is *not* perceived. Because in both perturbations (according to Fig. 5 and Fig. 19) the same points are left unaltered in the left and right fields and the same points are also perturbed, the fact that depth can be perceived in one case and not in the other removes the above objection.

Even this experiment is not a final proof that monocular micropattern recognition does not play some part in fusion. It might still be argued that this striking difference between depth perception, using for perturbation the unmixed grid (Fig. 5) or the mixed grid (Fig. 19), could be

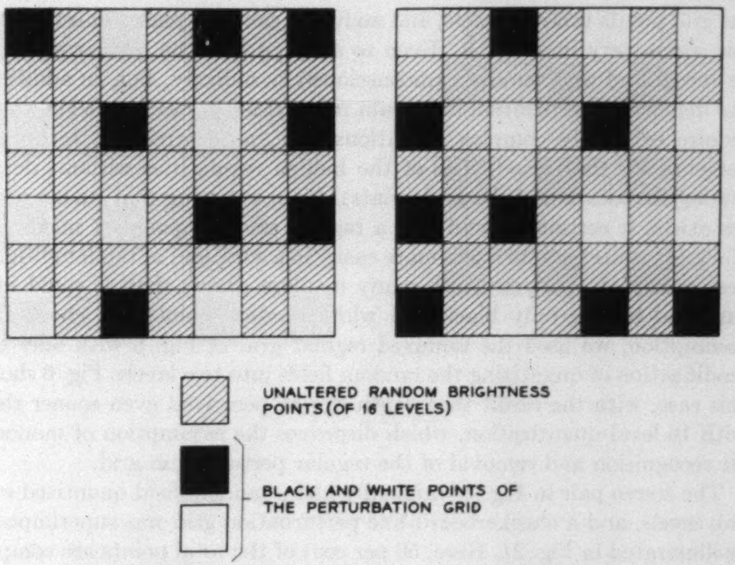

UNALTERED RANDOM BRIGHTNESS POINTS (OF 16 LEVELS)
BLACK AND WHITE POINTS OF THE PERTURBATION GRID

Fig. 18 — Illustration of the method by which the mixed regular perturbation grid was generated.

explained by this hypothesis of monocular pattern recognition: In the unmixed case the regular grid might be recognized monocularly by an unconscious process, then disregarded, and the remaining random points could now be fused monocularly without any difficulty. In the case of the mixed grid, this grid is not apparent monocularly, so the removal of

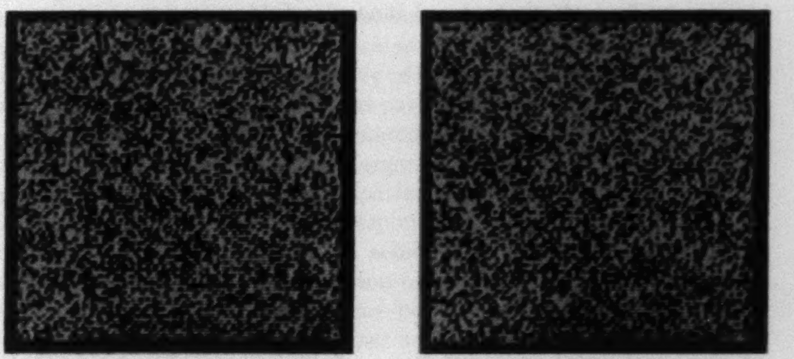


Fig. 19 — Stereo pair with mixed regular perturbation grid.

the grid points is not possible, and no fusion can take place. This hypothesis seems very improbable. Even to suppose that the regular grid can be recognized and removed unconsciously is unlikely, but, in addition, the monocular recognition of certain regularities in random fields would require extremely complex operations (e.g., autocorrelation technique detects only the periodicities of the hidden regularities without determining the location of the grid points). Even assuming that such a process exists, it certainly could find a regular grid composed of maximum black (or white) points much more easily in a surround of random brightness points of 16 levels (with many medium grays) than it could in a surround having only black and white random points. To check this assumption, we used the unmixed regular grid of Fig. 5 with only the modification of quantizing the random fields into two levels. Fig. 6 shows this case, with the result that depth can be perceived even sooner than with 16-level quantization, which disproves the assumption of monocular recognition and removal of the regular perturbation grid.

The stereo pair in Fig. 20 originally had a random field quantized into two levels, and a checkerboard-like perturbation grid was superimposed as illustrated in Fig. 21. Here, 50 per cent of the total points are complemented, and the regular grid has a double periodicity. Even in this case the depth effect can be easily obtained by fusing Fig. 20.

In these last experiments, the left and right images differ from each other considerably and the monocular recognition of the perturbation grid is made very difficult, yet we can still fuse the unaltered points with ease. These results disprove the hypothesis of monocular pattern recognition (both in the micro and macro sense), and suggest the second alternative: that the two fields are first combined and all further processings are performed on the fused binocular field.

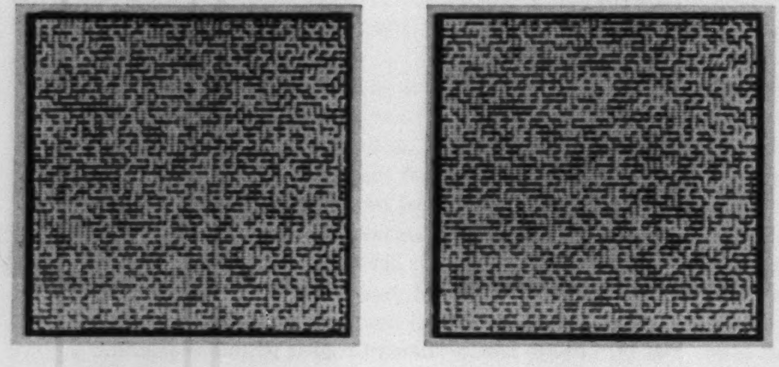


Fig. 20 — Stereo pair with "checkerboard" perturbation grid.

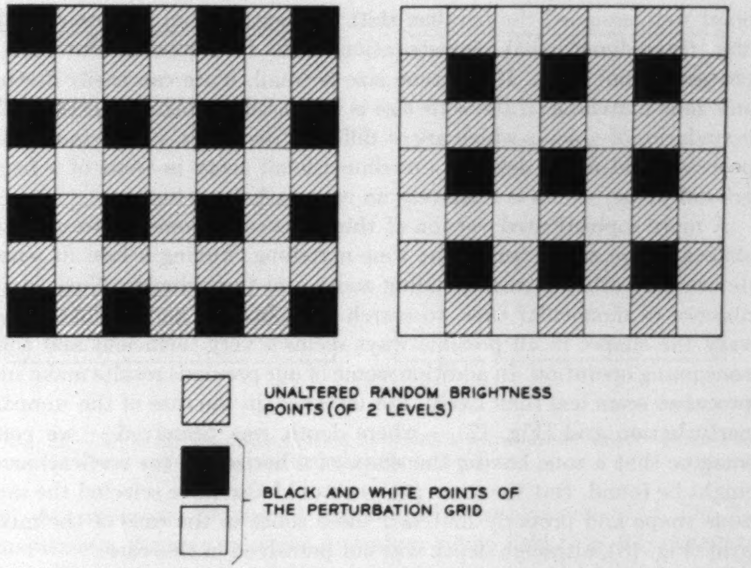

UNALTERED RANDOM BRIGHTNESS POINTS (OF 2 LEVELS)
BLACK AND WHITE POINTS OF THE PERTURBATION GRID

Fig. 21 — Illustration of the method by which the "checkerboard" perturbation grid was generated.

VIII. THE CONNECTION BETWEEN BINOCULAR PATTERN RECOGNITION AND DEPTH PERCEPTION

The demonstrations in the previous sections strongly suggest that, under these conditions, the perception of depth utilizes certain processings performed entirely on the fused binocular field. We intentionally do not yet call these processings binocular pattern recognition, because we must first investigate the feasibility of some processes that in ordinary usage are regarded as simpler than pattern recognition.

It has already been shown that matching corresponding point domains in the two fields does not require organizing these point domains into a higher entity of monocular macropatterns or micropatterns. One might think that the matching of corresponding point domains (instead of corresponding patterns) could be achieved by searching for a best fit according to some similarity criterion (e.g., maximum cross-correlation). A simple way to find correspondence between points in the two fields is to select a zone (of arbitrary shape) around any point in the left field and search for a zone in the right field (having the same shape) that is most similar to the left zone according to a given criterion. If this zone-matching were performed for every point in the visual field and each point were assigned the parallax shift (or depth value) so obtained, the final three-dimensional representation could be achieved. But such a process cannot work. If the zone size is small, noise can easily destroy any zone-matching; if the zone size is increased, ambiguities arise at the boundaries of objects which are at different distances. For instance, this process could never detect a one-dimensional array in front of a background plane, which is relatively an easy task for a human.

A more sophisticated version of this processing would be to vary the shapes of the zones during the zone-matching; finding a best fit would determine both the corresponding zones and their shapes. Now, in the absence of monocular cues, to search for a best fit and simultaneously vary the shapes in all possible ways seems a very inefficient and time-consuming operation. In addition, some of our previous results make such processes seem less than likely. For instance, in the case of the unmixed perturbation grid (Fig. 17) — where depth was perceived — we could imagine that a zone having the shape of a horizontal (or vertical) array might be found. But the same process would also have selected the same zone shape and properly matched these zones in the case of the mixed grid (Fig. 18), although depth was not perceived in this case.

Thus, it seems difficult to find simple operations (avoiding the use of pattern recognition) that give depth information consistent with that abstracted by the human visual mechanism. However, it is possible to demonstrate certain properties of point domains that are necessary in order for them to be seen in depth. These properties incorporate concepts such as connectivity, minimum size of a point domain, organization of close or periodic parts in higher entities, etc. We intuitively associate these notions with pattern-recognition operations. Therefore, our findings suggest that, under certain conditions, the perception of depth depends upon binocular pattern recognition. There is, of course, no evidence that this pattern recognition on the binocular field is identical to ordinary (monocular) pattern recognition. Nevertheless, an understanding of binocular pattern recognition may well be revealing when the broader aspects of pattern perception are considered. We will proceed, therefore, to investigate certain properties of patterns in the binocular field that yield depth effects.

The first question usually raised is this: Must the point domains possess any familiar pattern for them to be seen in depth? The answer is no. Any connected point domain can be seen in depth regardless of the shape of its boundary. The point domain should be connected at least in one dimension. This one-dimensional connectivity is a trivial property, which every object in real life possesses, and the following experiments show

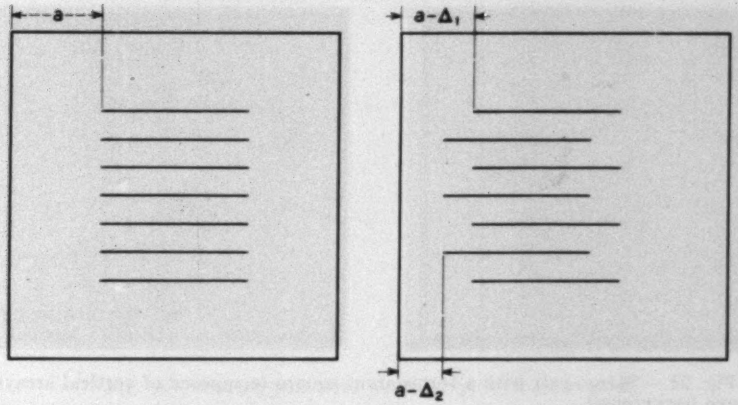


Fig. 22 — Illustration of the method by which a transparent center square was generated above another square (using horizontal arrays).

that this important property is preserved in the binocular field. Fig. 22 demonstrates the way in which a random stereo field (Fig. 23) is generated, with every even line (of 40-picture-element length) having a parallax shift of $\Delta_1 = 4$, and every odd line having one of $\Delta_2 = 6$.

The even and the odd lines each form a square that can be seen in depth; the far one appears to have a regular surface; the closer square seems transparent. Either horizontal or vertical connectivity yields the same results. Fig. 24 shows such a case, where the pattern is composed of vertical random arrays of 40 picture elements in length. Twenty even

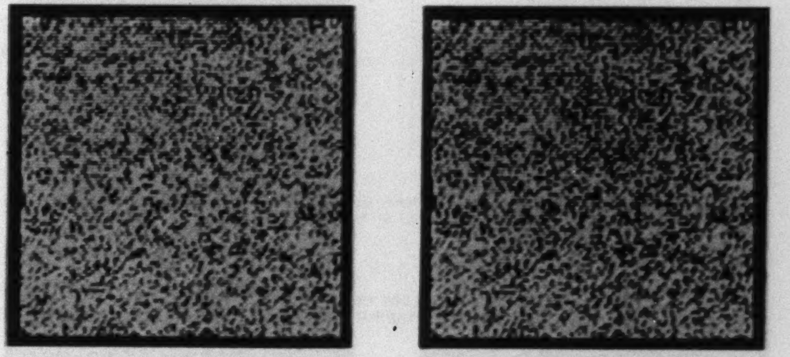


Fig. 23 — Stereo pair with a transparent square (composed of horizontal arrays) above the center square.

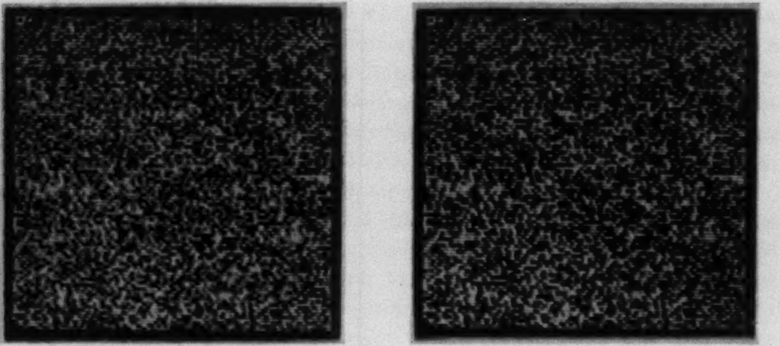


Fig. 24 — Stereo pair with a transparent square (composed of vertical arrays) above background.

vertical arrays form the "transparent" center square; the odd vertical arrays belong to the background.

If we now try an experiment using isolated points of the same depth, it is very difficult to see these points forming a ghost-like plane, even if the points are regularly spaced. Fig. 25 shows such a case, where the regular presentation of Fig. 4 is used but every second point in every second line has a parallax of $\Delta_2 = 2$. If these isolated points at the same distance are not regularly spaced and not dense enough, they cannot be organized as forming one surface.

To show the importance of connectivity in another example, Fig. 26 demonstrates a stereo pair with a meshlike perturbation grid (shown in Fig. 27). Although 50 per cent of the points are unaltered (as in Fig. 20),

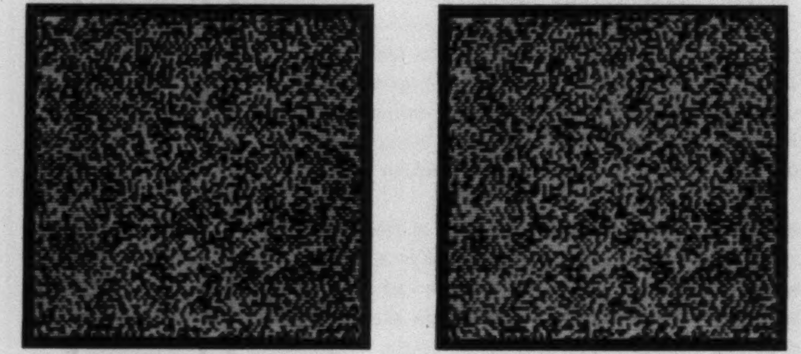


Fig. 25 — Stereo pair with "ghost" square (composed of isolated points) above background.

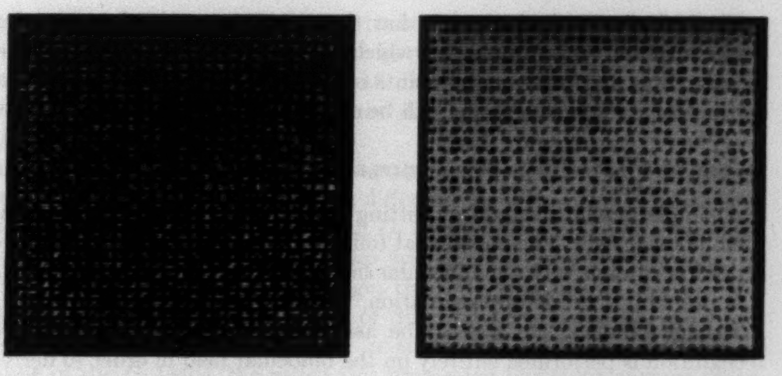


Fig. 26 — Stereo pair with meshlike perturbation grid.

the depth effect is now greatly reduced. The only explanation offhand is the fact that the perturbation mesh limits the connectivity to small, separated subdomains. It is also interesting that these subdomains must possess a critical size in order to be seen in depth. The investigation of this quantitative aspect is not attempted at the present.

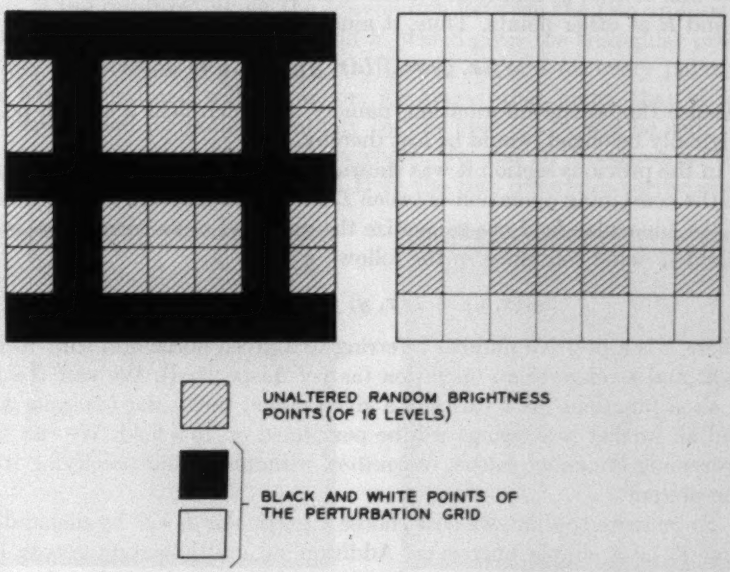

UNALTERED RANDOM BRIGHTNESS POINTS (OF 16 LEVELS)
BLACK AND WHITE POINTS OF THE PERTURBATION GRID

Fig. 27 — Illustration of the method by which a meshlike perturbation grid was generated.

These findings might suggest that the patterns seen in the binocular field are similar to contour lines, which consist of continuous one-dimensional arrays and connect the points of equal parallax shift. In the next section a simple analog model will be derived along these lines.

IX. THE DIFFERENCE FIELD AS A SIMPLE ANALOG TO THE BINOCULAR FIELD

A simple model is an aid in getting greater insight into properties of the binocular field. The model that follows appears to have several properties in common with the binocular field as perceived, but on the whole it is probably a crude approximation.

In the following we accept the assumption that binocular pattern recognition is performed entirely on the binocular field in order to derive depth information, and we remember that the image points belonging to the left and right fields must be labeled. The binocular field $f(L, R)$ is a function of the left and right fields (L and R); thus, the set of all points in the binocular field is a function of the set of brightness points $L(x, y)$ and $R(x, y)$ in the monocular fields, where x and y are the coordinates.

Now the value of $f(L, R)$ at some point x, y must not be merely a function of $L(x, y)$ and $R(x, y)$, but must in fact depend on the values of L and R at other points. Thus, it must not be of the form

$$f(x, y) = f[L(x, y), R(x, y)]$$

because the crucial information, namely, to which field a certain point originally belonged, would be lost thereby.

In the previous section it was shown that cross-correlation also cannot be the combining operation between L and R. To derive a simple model of the binocular field, we generalize the notion of cross-correlation, with L and R being combined in the following way:

$$g_k(x, y) = L(x, y) * R(x + k, y),$$

where k is a positive number referring to a given horizontal shift to the right and $*$ refers to an operation (as yet unspecified). We call the set of all g_k functions (as k varies in a given range) the *analog binocular field* and all further processings will be performed on this field. We call this processing *binocular pattern recognition* without further specifying it at the present.

To be more specific, we now choose a particular $L * R$ by demanding that it be a simple operation. Addition or multiplication seems less favorable than substraction or division; this assumption is based on the experiments with Fig. 17, where the perturbation with an unmixed grid gave depth effect, and Fig. 18, where the mixed regular grid did not. Neither $g_0 = L(x, y) + R(x, y)$ nor $g_0 = L(x, y) \cdot R(x, y)$ would discriminate between Fig. 17 and Fig. 18 (being identical for both cases), whereas both $L(x, y) - R(x, y)$ and $L(x, y)/R(x, y)$ could account for the difference in depth impression.

Finally, we choose $D_k(x, y) = L(x, y) - R(x + k, y)$ as the simplest operation at hand, and call D_k the difference field having a parallax shift of k picture elements. The set of all D_k fields is an analog binocular field, which is designated as the *difference field* D. In these investigations, we limit k to integers in a given range; thus, the final model consists of a finite number of difference fields of different parallax shifts. Now, determining the binocular parallax is equivalent to finding patterns in some of the D_k fields. We called this processing binocular pattern recognition, and in this analogy we regard it as being identical to *ordinary (monocular) pattern recognition.*

In the case of our regular presentation (that is, the random stereo field containing a square with a parallax shift of four picture elements surrounded by a background with zero parallax shift), the following difference fields will be obtained: (a) D_k for $k \neq 0$ or 4 are random fields where each brightness point has a triangular probability distribution [this is the result of taking the convolution between the two uniformly distributed random variables L and R, which gives the triangular probability distribution of $(L - R)$]; (b) D_0 will be zero for every point in the background and will be random for every other point, that is, for the square and for points seen only by one eye; (c) D_4 will be zero for the central square and random elsewhere. (D_0 and D_4 are shown as the left and right pictures in Fig. 28.) Here the zero difference corresponds

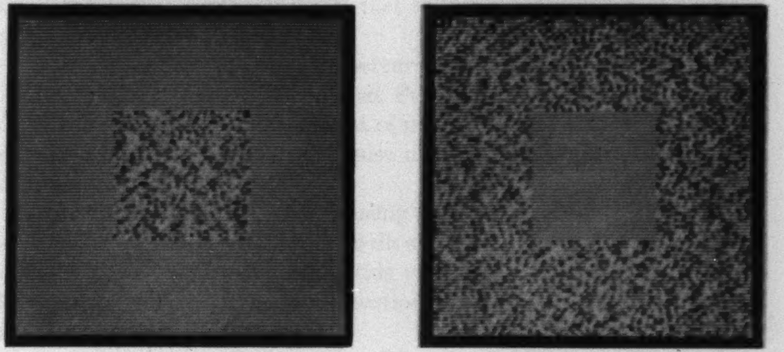


Fig. 28 — Difference fields D_0 and D_4 for the case of Fig. 4.

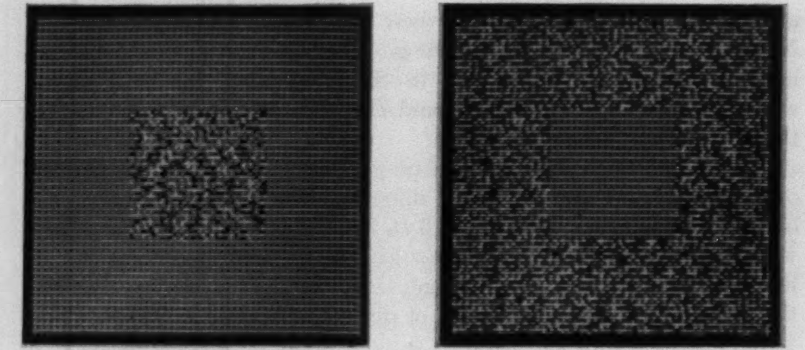


Fig. 29 — Difference fields D_0 and D_4 for the case of Fig. 5.

to a medium-gray level, the maximum positive difference to maximum white and the maximum negative difference to maximum black.

Only D_0 and D_4 are presented, because all other difference fields consist entirely of random brightness points. In the case of familiar stereo pairs, the difference field D_k contains points of near-zero value forming contour-lines having equal parallax shifts of k picture elements.

The next two pictures, in Fig. 29, show D_0 and D_4 for the case of the unmixed perturbation grid in Fig. 5. Through the perturbation grid, the uniformly gray background (or square, respectively) is clearly visible. Now, taking the mixed perturbation grid in Fig. 19, D_0 and D_4 should be very similar to the unmixed case. In the unmixed case the perturbation grid is always black (or always white) for D_0, which for the mixed case

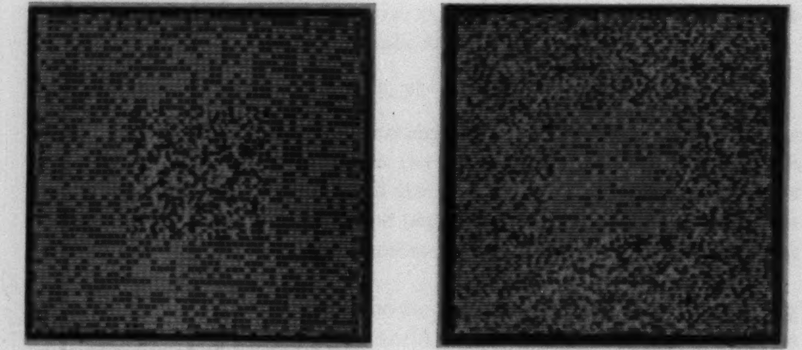


Fig. 30 — Difference fields D_0 and D_4 for the case of Fig. 19.

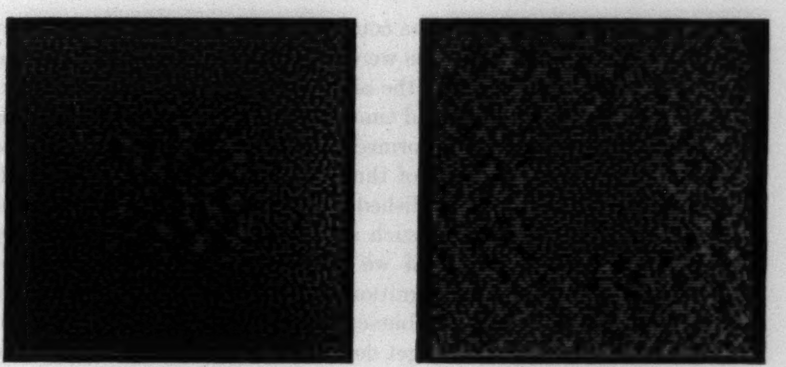


Fig. 31 — Difference fields D_0 and D_4 for the case of Fig. 13.

yields the same regular grid, but the grid points can take black and white at random. The left picture in Fig. 30 shows D_0 in this case, and it is now striking how well, in contrast to Fig. 29, the random central square is hidden by this type of perturbation. The right picture in Fig. 30 is D_4 for the mixed perturbation grid. Here, the grid points can take black and white values with 25 per cent probability each, and gray values with 50 per cent probability. Therefore, only 12.5 per cent of D_4 is effectively perturbed, but, because of the random appearance of this perturbation, it is more effective in hiding the central square than is 25 per cent perturbation of the unmixed grid. The uniform regions must be detected both in D_0 and D_4 to get depth.

In the next picture (Fig. 31), D_0 and D_4 are presented for the blurred picture of Fig. 13. The separation between the square and background is clearly visible, which confirms the fact that depth is also well perceived in this case.

By introducing gaussian noise perturbation in the stereo pairs (as in Fig. 11), D_0 and D_4 were determined. Subjective experiments were then conducted to determine the amount of noise that cancels depth, and this amount was compared with the noise needed to hide the square in the difference fields.

The results of this experiment, using ten subjects, indicated that the threshold of perceiving depth was 6 db signal to noise (with a very rapid decline in depth perception below this value), and that the *same* threshold value was obtained for the detection of the square in the difference fields.

As was emphasized before, the difference fields are probably very crude analogies for the binocular fields; nevertheless, it is worthwhile to mention the following fact: In the course of these investigations a great number of different perturbations were introduced in the stereo random fields. As a result of this process, the obtained stereo pairs could be rank-ordered according to stability and time required to perceive depth. This same ordering process was performed on the corresponding difference fields based on the separability of the central square and its surround. It turned out that the two established hierarchies were identical, except for borderline cases. Naturally, such subtleties cannot be explained by our simple analogy, especially if we consider the following: We performed monocular pattern recognition on the difference field in order to detect certain regions, while binocular pattern recognition was performed on the binocular field to get depth. There is no evidence that the laws of binocular pattern recognition are identical to ordinary (monocular) pattern recognition. (For instance, it is known that connectivity is an important monocular pattern-recognition cue that seems to be even more emphasized in binocular pattern recognition.)

Even the assumption of using a linear operation (subtraction) in the model is naturally an oversimplification. In the next experiment we demonstrate a nonlinear phenomenon of the binocular space. The perturbation grid in Fig. 32 is used. Here, every even sample in every even

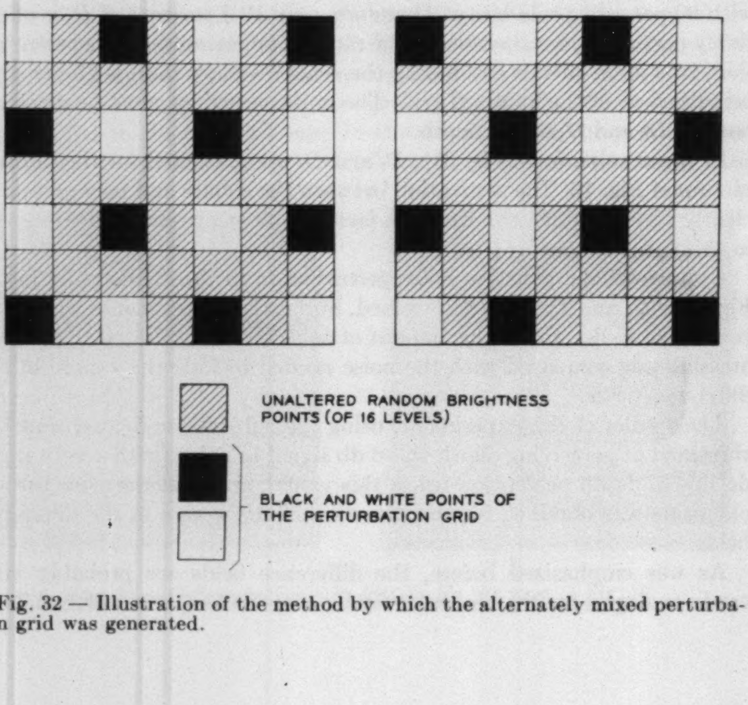

UNALTERED RANDOM BRIGHTNESS POINTS (OF 16 LEVELS)
BLACK AND WHITE POINTS OF THE PERTURBATION GRID

Fig. 32 — Illustration of the method by which the alternately mixed perturbation grid was generated.

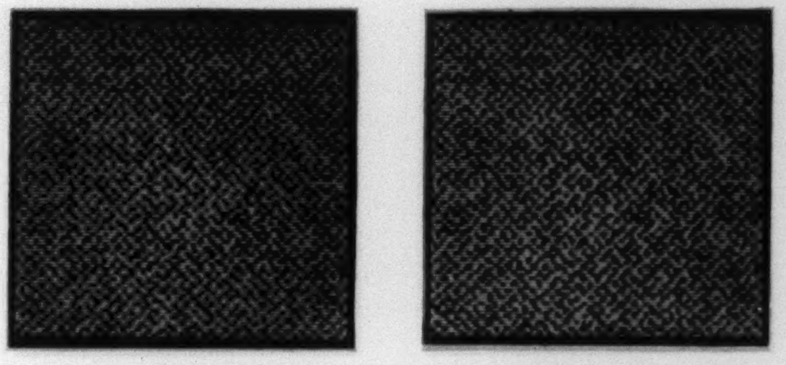


Fig. 33 — Stereo pair with alternately mixed perturbation grid.

line is alternatingly black and white, and its complemented value is copied in the other stereo picture. Fig. 33 demonstrates this case. Under strong illumination depth cannot be perceived. When the lights are dimmed (or eyes squinted), depth is easily obtained. Fig. 34 shows the difference fields; here, we find that detection of the center square is somewhat dependent on the illumination. However, this weak dependence is not consistent with the depth experiment.

X. MONOCULAR MACROPATTERNS ENHANCE DEPTH PERCEPTION

In posing our original problem we were interested in whether the perception of depth uses monocular pattern recognition, binocular pattern recognition or a combination of both. In the previous sections it was demonstrated that depth can be perceived without monocular patterns

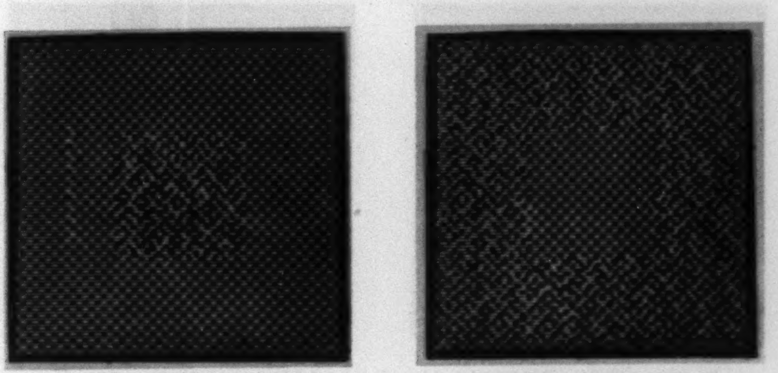


Fig. 34 — Difference fields D_0 and D_4 for the case of Fig. 33.

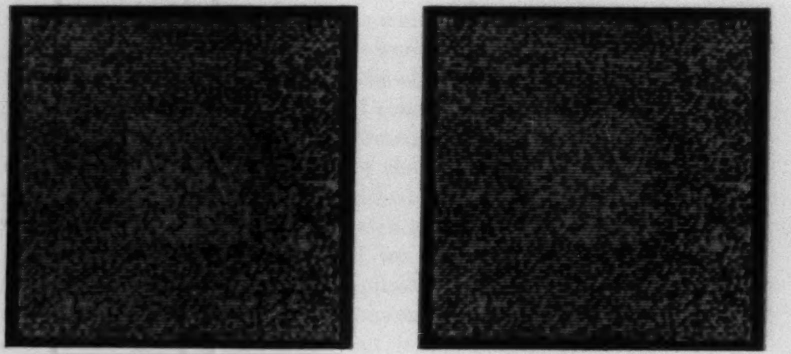


Fig. 35 — Stereo pair with brighter center square.

being present. In this section it will be demonstrated, nonetheless, that monocular macropattern recognition enhances depth perception. The same random stereo images are used, but the average value of the brightness points of the square is increased. Because of this, the random points in the square are brighter than the surround, and the square can be also seen monocularly. Fig. 35 demonstrates this case; it is apparent that the depth effect is obtained much faster than it is with missing monocular cues. According to this, we can suppose that depth perception is a combination of binocular and monocular pattern recognition, as was suggested in Fig. 3.

The actual processes of depth perception are, of course, much more complicated than the simplified diagram in Fig. 3. The different blocks are probably connected in many ways. Complicated feedback loops exist

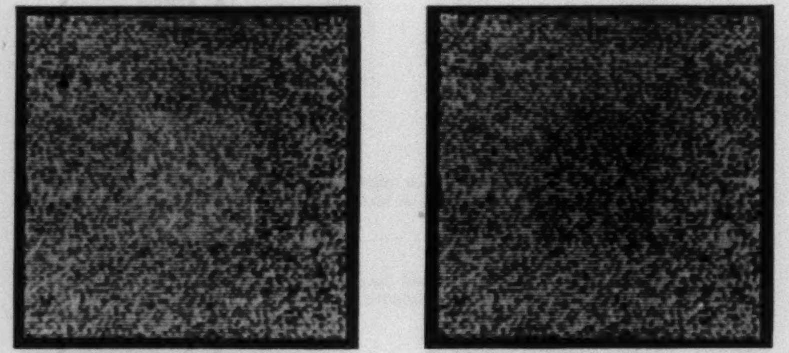


Fig. 36 — Stereo pair with whiter left and blacker right center square.

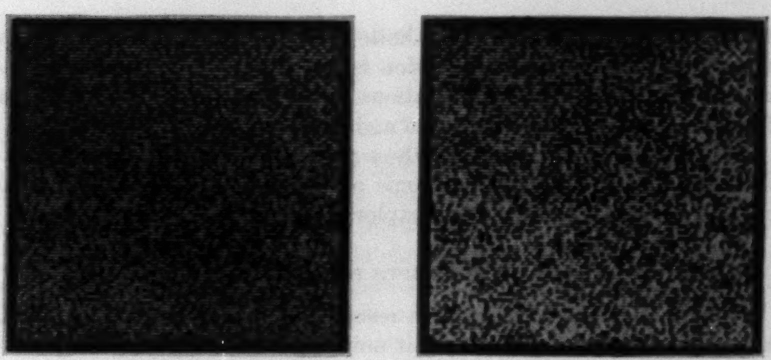


Fig. 37 — Stereo pair with left picture attenuated three times.

between the binocular field and the monocular fields, between the binocular pattern recognizer and the depth perceiver, etc. Fig. 36 demonstrates such a feedback between the binocular and monocular fields. Here, the random points in the left square have a mean value 20 per cent less than the surround and 20 per cent more than the surround in the right field. By fusing Fig. 36 we can see the square in depth with apparently the same brightness as the surround.

Fig. 37 shows another interesting case, where the left brightness values are attenuated by dividing them with a factor of three. In this experiment, $\Delta = 7$ picture elements and the center square is only 30×30. Depth is still easily perceived, according to expectation.[7]

Another even more complicated operation takes place in the monocular fields in connection with the binocular field. In Fig. 38 the left pic-

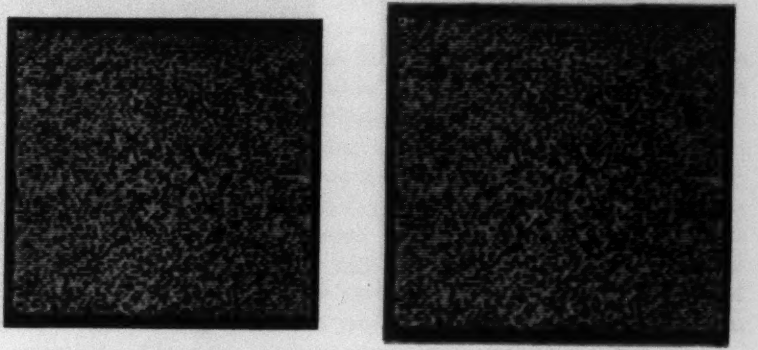


Fig. 38 — Stereo pair with right picture expanded by 10 per cent.

ture is contracted 10 per cent in both height and width. Even with this tremendous size discrepancy, fusion is possible and depth can be perceived. The same is true for rotations. More than ± 6 degrees rotation from the base line can be tolerated and depth perceived.

The thorough investigation of these processes is the key to real understanding of depth perception. Some of the techniques developed here might be useful in such further exploration.

XI. SOME PROPERTIES OF THE DEPTH PERCEIVER

In Figs. 1, 2 and 3 the pattern recognizers were followed by a block called the "depth perceiver." This unit might have the function of coordinating several pattern-recognition tasks and assigning depth to various points. Even those points that have no parallax (seen by one eye only) will be located in depth. When there is no contextual reason to assign a particular depth to certain ambiguous point domains, there is a general tendency to see them in the farthest plane.

This tendency can be demonstrated by fusing Fig. 39. Here, the ambiguous random points lie in the place of the uniform black square seen behind the surround. Some investigations of ambiguous stereo effects (without parallax shift) were recently carried on with a similar result.[8]

The depth perceiver is particularly sensitive to any vertical shift (perpendicular to the base line). Parallax shifts with slight vertical components will not give rise to depth effects, probably because such shifts cannot occur in life. It seems reasonable to assume that the depth perceiver utilizes monocular depth cues too.

Naturally, all such divisions into different blocks are mere specula-

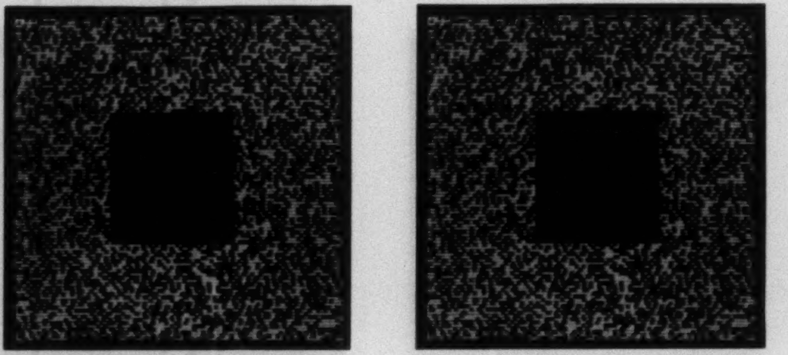


Fig. 39 — Stereo pair with uniformly black center square behind the random foreground.

tions until other psychological and physiological findings give adequate support.

XII. CONCLUSION

The peculiar depth effects that have been demonstrated strongly suggest that, under these conditions, depth perception is closely related to pattern recognition processes on the binocular field. Someone could raise the question: What is the merit of showing that binocular and not monocular pattern recognition is required in depth perception if the processes of pattern recognition are still unknown?

To answer this, we must realize that pattern-recognition processes are complex and highly nonlinear in nature. Because of this, it is very important which operations are performed on the input patterns before recognition. (For instance, upon performing the pattern-recognition task on the difference fields of Fig. 29 and Fig. 30, the qualitative difference of perceiving depth in the two cases is instantly apparent, which could not be simply explained if the recognition had been performed on the monocular patterns of Fig. 5 and Fig. 19.)

Thus, the discovery of certain transformations of the input patterns that facilitate the recognition task provides better understanding of the laws of pattern recognition.

These experiments indicated also that, without monocular cues or *Gestalt*, depth can be still perceived. In order to be seen in depth, the patterns need to possess much simpler properties (e.g., one-dimensional connectivity, adequate number of connected points, etc.) than we originally expected. These properties might be simple enough to be simulated by present computer technology. Thus, the findings of this study might give a new impetus to the development of devices that will determine depth automatically.

The technique of stereo random fields also has several advantages in a great variety of possible applications. In binocular fusion studies, the problem of binocular rivalry sometimes makes investigation cumbersome. These stimuli have a self-checking feature against binocular rivalry; namely, as long as depth is seen, no rivalry can be present.

The long time constants needed to perceive depth in certain presentations indicate that depth perception depends very much on the input material. From the order of a few milliseconds (required for simple stereo pictures), we can easily increase the perception time to the order of minutes. This slowing down of a process can be very advantageous in investigations of learning, pattern recognition, etc.

The stability of the random stereo fields is also very useful. Because

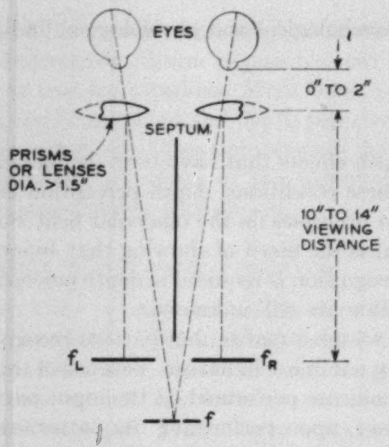

EYES
0" TO 2"
SEPTUM
PRISMS OR LENSES DIA. > 1.5"
10" TO 14" VIEWING DISTANCE
f_L
f_R
f

Fig. 40 — Illustration showing how presented stereo pictures should be viewed.

nearly all points carry depth information, the stereo image is very stable and points with greater parallax shifts than in the ordinary case can be fused.

Such stimuli could also possibly be used in apparent motion studies.

This technique was found to be a useful tool in color studies to examine the role of color in depth perception.

But perhaps the most useful property of this method is the elimination of context and higher organization from the input stimulus, which makes it possible to isolate and study less formidable problems.

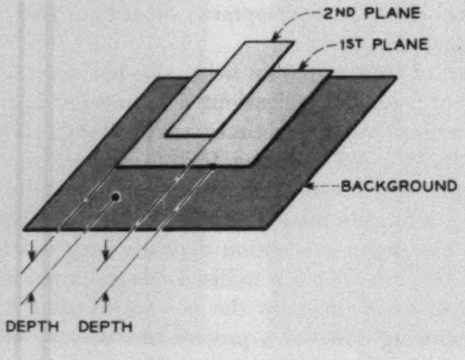

2ND PLANE
1ST PLANE
BACKGROUND
DEPTH
DEPTH

Fig. 41 — The subjective illusion seen when Fig. 8 is viewed stereoscopically.

XIII. ACKNOWLEDGMENT

I am indebted to E. E. David, Jr. for his valuable comments on this paper and to R. A. Payne for his skillful assistance in taking and processing the many photographs.

APPENDIX

The presented stereo pictures can be fused if they are viewed through a pair of lenses used as prisms, as shown in Fig. 40. The focal length of the lenses should be 10 to 18 inches and their diameter around 1½ inches or more, as is the case with the ones accompanying this paper. Sometimes it takes several minutes to get the depth effect.

If fusion of the left and right images cannot be obtained easily, a stiff paper or cardboard septum (10 to 14 inches long) placed between the two stereo pictures and perpendicular to the page will probably eliminate the difficulty (see Fig. 40). Viewers who ordinarily wear glasses should not remove them when using the lenses.

For example, the subjective illusion that is seen when Fig. 8 is viewed stereoscopically is illustrated in Fig. 41.

Lenses for use with "Binocular Depth Perception of Computer-Generated Patterns" by Bela Julesz in September 1960 Bell System Technical Journal, pp. 1125-1161.

These lenses should be used as illustrated in Fig. 40 on page 1160. If fusion of the left and right images cannot be obtained easily, a stiff paper or cardboard septum (10 to 14 inches long) placed between the two stereo pictures and perpendicular to the page will probably eliminate the difficulty. Viewers who ordinarily wear glasses should not remove them when using the lenses.

REFERENCES

1. Gibson, J. J., Perception of Distance and Space in the Open Air, Army Air Force Program, Report No. 7, 1946, p. 181; reprinted in Beardslee, D. C. and Wertheimer, M., eds., *Readings in Perception*, D. Van Nostrand Co., New York, 1958.
2. Langlands, H. M. S., Experiments in Binocular Vision, Trans. Opt. Soc. London, **28,** 1926, p. 45.
3. Wheatstone, C., On Some Remarkable, and Hitherto Unobserved, Phenomena of Binocular Vision, Roy. Soc. London Phil. Trans, 1838, p. 371.
4. David, E. E., Jr., Mathews, M. V. and McDonald, H. S., Experiments with Speech Using Digital Computer Simulation, I.R.E. Wescon Conv. Rec. — Audio, August 1958, p. 3.
5. Graham, R. E. and Kelly, J. L., Jr., A Computer Simulation Chain for Research on Picture Coding, I.R.E. Wescon Conv. Rec.—Computer Applications, August 1958, p. 41.
6. Julesz, B., A Method of Coding Television Signals Based on Edge Detection, B.S.T.J., **38,** 1959, p. 1001.
7. Ogle, K. N. and Groch, J., Stereopsis and Unequal Luminosities of the Images in Two Eyes, A.M.A. Arch. Ophthal., **56,** 1956, p. 878.
8. Wegner, K., Stereoskopische Untersuchungen für Tiefenlokalisation sogenannter funktionsloser Bestandteile des Binocularen Gesichtsfeldes, Thesis, Georg-August Univ., Göttingen, Germany, 1959.

Models for Approximating Basilar Membrane Displacement

By J. L. FLANAGAN

(Manuscript received April 1, 1960)

Three analytical models are developed for estimating the displacement of the basilar membrane in the human ear when the sound pressure at the eardrum is known. Frequency-domain data, derived experimentally by Bekesy, are Fourier-transformed to examine the impulse response of the membrane. Time-domain and frequency-domain responses of the models are compared with the experimental data. Excitation of the models by periodic impulses is considered. Calculations of membrane displacement are made for excitation by positive pulses, and by alternately positive and negative pulses. Applicability of the results to the perception of pitch is indicated.

I. INTRODUCTION

In the course of developing an hypothesis to account for results obtained in two experiments on pitch perception,[1,2] it became desirable to have a tractable model from which the displacement of the basilar membrane at a given point could be estimated from a knowledge of the sound pressure at the eardrum. This report describes the results of an effort to deduce such a model.

II. MECHANICAL PROPERTIES OF THE MIDDLE EAR AND COCHLEA

To recall facts and establish a frame of reference, a simplified sketch of the peripheral mechanism of hearing is shown in Fig. 1. The cochlea, actually wound in a snail-shell-like spiral in man, is sketched here unrolled and stretched out. It contains the perilymph fluid and is partitioned longitudinally by a duct formed by Reissner's membrane and the basilar membrane. The duct, roughly triangular in cross section, is filled with another fluid, endolymph. Resting upon the basilar membrane within the cochlea duct is the organ of Corti. This organ, immersed in the endolymph, serves as the termination of the auditory nerve. Bekesy[3] has established that the basilar membrane and Reissner's membrane

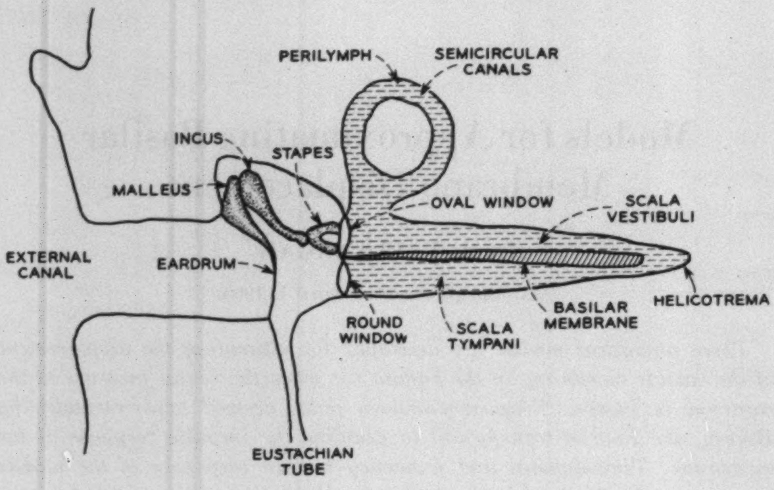

PERILYMPH
SEMICIRCULAR CANALS
INCUS
STAPES
MALLEUS
OVAL WINDOW
SCALA VESTIBULI
EXTERNAL CANAL
EARDRUM
ROUND WINDOW
SCALA TYMPANI
BASILAR MEMBRANE
HELICOTREMA
EUSTACHIAN TUBE

Fig. 1 — Schematic drawing of the human ear.

vibrate cophasically when the ear is stimulated by sound in the lower range of audible frequencies. Because Reissner's membrane does not enter into the present development, only the basilar membrane is sketched in the schematic diagram.

A sound wave impinging on the ear is led down the external canal and sets the drum into vibration. The vibration is transmitted by the ossicular chain to the cochlea, where the piston-action of the stapes foot-plate produces a compressional wave in the fluid. Because of its distributed mass and elastic and viscous constants, and because of the pressure release at the round window, the basilar membrane vibrates selectively according to the frequency content of the stimulus. Displacement of the basilar membrane causes pressure to be exerted (by another membrane in the cochlea duct, the tectorial) upon the hairs emanating from hair cells in the organ of Corti. When the hairs are sufficiently deformed, electrical discharges are produced in the nerve fibers.

The mechanical properties of the cochlea have been studied in detail by Bekesy.[4] He found that, when the stapes is driven sinusoidally with constant amplitude of displacement, the amplitude of displacement of points along the low-frequency (or apical) end of the basilar membrane varies with frequency as shown in Fig. 2. The peak displacement of each point is normalized to unity. His measurements[3] of the difference in phase between the displacement of the stapes and the displacement of

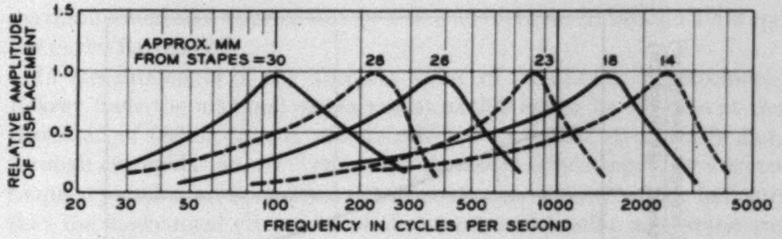


Fig. 2 — Relative amplitude of displacement as a function of frequency for different points along the basilar membrane. The stapes is driven with constant amplitude of displacement (after Bekesy[4]).

points along the membrane are sketched in Fig. 3. In addition to these data, Bekesy found[5] that, when the sound pressure is constant at the eardrum, the magnitude of volume displacement of the round window is nearly constant up to around 2000 cps. To the extent that the perilymph is incompressible and the walls of the cochlea rigid, the volume displacement of the round window is equal that of the stapes footplate.

Data reported by Zwislocki[6] and by Bekesy[5] indicate that, for frequencies below 1000 cps, the over-all impedance of the middle ear and cochlea is predominantly elastic, owing principally to the compliance of

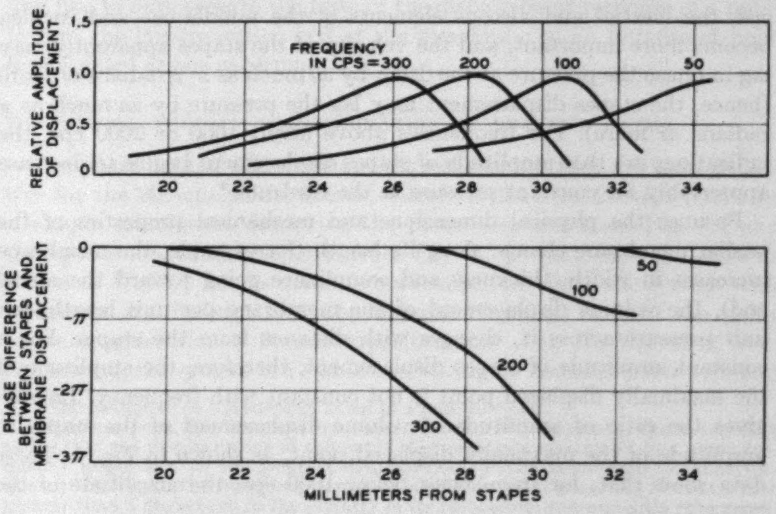


Fig. 3 — Relative amplitude and phase of basilar membrane displacement as a function of distance along the membrane (after Bekesy[3]).

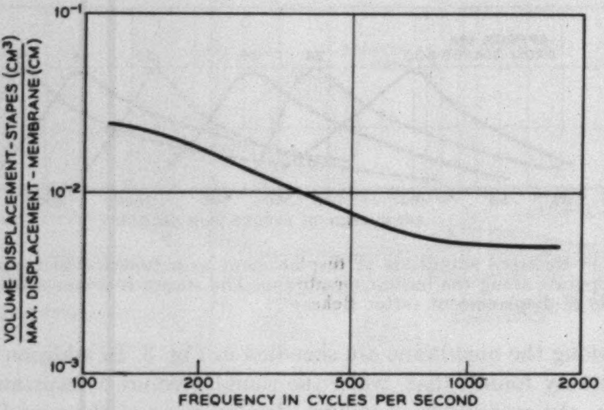


Fig. 4 — Ratio of volume displacement of stapes to peak displacement of basilar membrane (after Bekesy[4]).

the middle ear air cavity, the round window membrane and the ligaments retaining the ossicles and drum. For these frequencies, therefore, the displacement of the stapes is essentially proportional to, and in phase with, the sound pressure at the eardrum. At frequencies above 1000 cps, the inertial and viscous elements of the middle ear and cochlea become more important, and the velocity of the stapes apparently may lag in phase the pressure at the drum by as much as $\pi/2$ radians or more (hence, the stapes displacement may lag the pressure by as much as π radians or more). For frequencies above about 1000 or 2000 cps, the indications are that amplitude of stapes displacement begins to decrease appreciably for constant pressure at the eardrum.[*]

Because the physical dimensions and mechanical properties of the basilar membrane change along its length (for example, the membrane increases in width, thickness and compliance going toward the apical end), the volume displacement of the membrane per unit length, per unit pressure across it, changes with distance from the stapes. For a constant amplitude of stapes displacement, therefore, the amplitude of the maximally displaced point is not constant with frequency. Bekesy[4] gives the ratio of amplitude of volume displacement of the stapes to amplitude of the maximally displaced point, as shown in Fig. 4. These data show that, for frequencies below 1000 cps, the amplitude of the maximum increases approximately 4 or 5 db/octave. At around 1000 cps the curve flattens off.

---

[*] Zwislocki's data suggest a decrease of the order of 12 to 18 db/octave; Bekesy's average data seem to agree roughly with this. In one preparation, however, Bekesy obtained a fall of about 30 db/octave.

In measurements of the absolute value of membrane displacement, Bekesy finds the maximal displacement at 200 cps to be 10^{-4} cm at the threshold of feeling (about 140 db referred to 0.0002 dyne/cm^2) and, through extrapolation, 10^{-11} cm at the threshold of hearing.* For a given frequency and a given point on the membrane, Bekesy's data indicate that the mechanical vibrations of the stapes and basilar membrane are essentially linearly related until sound pressures above the threshold of feeling are reached. There is evidence, however, that the ear is capable of producing perceptible subjective components at sound levels less than this value.

As stated at the outset, we desire an analytical relation for estimating the basilar membrane displacement at a given point from a knowledge of the sound pressure at the eardrum, valid at least in the frequency range below 1000 cps. It is in this range that the stapes displacement is in phase with, and proportional to, the pressure at the drum. The experimental data that the model must describe are the frequency-domain data just discussed. The approximation problem may, of course, be approached in either the time or frequency domains; usually it is helpful to maintain some insight in both domains. Consequently, we would first like to inquire as to the form of the displacement response of a point toward the low-frequency end of the membrane to an impulse of pressure applied at the eardrum.

III. INVERSE FOURIER TRANSFORMATION OF BEKESY'S DATA

The phase data of Fig. 3 are at best meager, but they are most definitive for the 200-cps point. Let us, therefore, take the 200-cps point for a sample calculation. Deducing the phase response from Fig. 3,† and taking the amplitude response from Fig. 2, we may plot the data as shown in Fig. 5.‡ Let us make two assumptions about the system with which we are dealing: first, the impulse response, $h(t)$, of the point under consideration is Fourier transformable (i.e., $\int_{-\infty}^{\infty} h^2(t)\,dt < \infty$); and second, the system is a stable one having no complex poles with real

---

* The diameter of a hydrogen atom is about 10^{-8} cm.

† Because peak displacement increases at around 5 db/octave, the possibility exists that the displacement of the point that responds maximally to a given frequency might not be the greatest displacement of the membrane for that frequency. However, the frequency response of a given point generally rises at a rate greater than 5 db/octave in the vicinity of its resonance; consequently, the greatest displacement occurs essentially at the maximally responding point.

‡ As closely as I can determine from the Akustische Zeitschrift data, the maximum displacement of the "200-cps point" falls at about 210–220 cps.

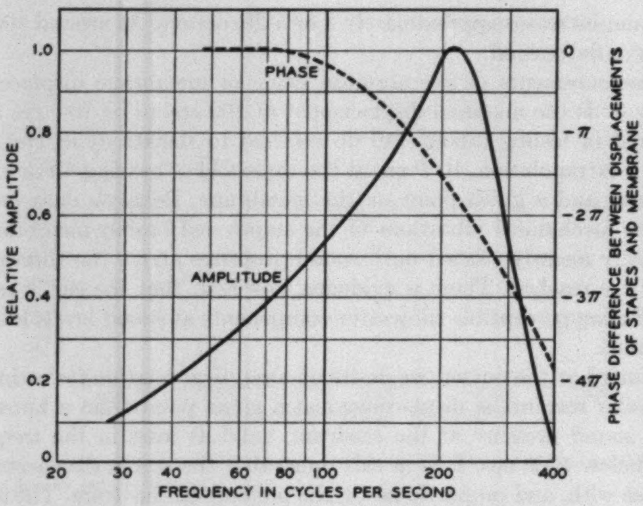


Fig. 5 — Displacement amplitude and phase for a point near the apical end of the basilar membrane. Maximum response occurs for a frequency of about 200 cps. These curves are obtained from data in Figs. 2 and 3.

parts equal to, or greater than, zero (i.e., the system exhibits no output until an input is applied, and the final value of the impulse response is zero).

Taking the data of Fig. 5 as the magnitude, $| H(\omega) |$, and phase, $\Phi(\omega)$, respectively, of the Fourier transform, $H(\omega)$, of the impulse response, $h(t)$, we wish to calculate the inverse transform:

$$h(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(\omega) e^{j\omega t} \, d\omega. \tag{1}$$

In Cartesian form, $H(\omega)$ is

$$H(\omega) = \mathrm{Re}\, H(\omega) + j \,\mathrm{Im}\, H(\omega),$$

where

$$\begin{aligned} \mathrm{Re}\, H(\omega) &= | H(\omega) | \cos \Phi(\omega), \\ \mathrm{Im}\, H(\omega) &= | H(\omega) | \sin \Phi(\omega). \end{aligned} \tag{2}$$

Because Re $H(\omega)$ is an even function of ω and Im $H(\omega)$ an odd function, (1) reduces to:

$$\begin{aligned} h(t) &= \frac{1}{\pi} \int_0^{\infty} \mathrm{Re}\, H(\omega) \cos \omega t \, d\omega - \frac{1}{\pi} \int_0^{\infty} \mathrm{Im}\, H(\omega) \sin \omega t \, d\omega \\ &= h_1(t) + h_2(t), \end{aligned} \tag{3}$$

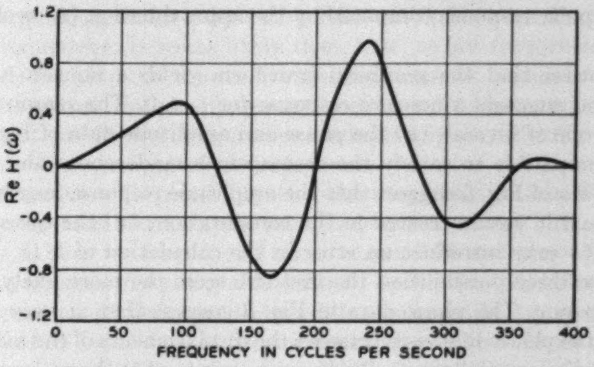


Fig. 6 — Real part of the Fourier transform, $H(\omega)$, whose amplitude and phase spectra are given in Fig. 5.

where $h_1(t)$ is an even function of time and $h_2(t)$ an odd function. Because of the assumptions regarding stability [i.e., $h(t) = 0$, for $t < 0$]:

$$h_1(t) = -h_2(t) \qquad \text{for} \qquad t < 0,$$

and

$$h_1(t) = h_2(t) \qquad \text{for} \qquad t > 0. \tag{4}$$

Hence (3) can be written:

$$h(t) = \frac{2}{\pi}\int_0^\infty \mathrm{Re}\, H(\omega) \cos \omega t \, d\omega \qquad \text{for} \qquad t > 0. \tag{5}$$

To calculate $h(t)$, then, only Re $H(\omega)$ is needed. For the data of Fig. 5, Re $H(\omega)$ is plotted in Fig. 6.*

In the absence of an analytical specification of Re $H(\omega)$, we have graphically evaluated the integral (5) by using the approximation:

$$h(t_i) = \frac{2}{\pi}\sum_{n=0}^{40} \mathrm{Re}\, H(\omega_n) \cos \omega_n t_i \Delta\omega, \tag{6}$$

where:

$$\begin{aligned} \omega_n &= n\omega_0\,, \\ \omega_0 &= (2\pi)(10) \text{ radians per second} \\ \Delta\omega &= (2\pi)(10) \text{ radians per second,} \\ t_i &= (0.4 \times 10^{-3})i, \qquad i = 0, 1, 2, \cdots, 27. \end{aligned}$$

---

* Re $H(\omega)$ was obtained from a large linear plot of $|\, H(\omega) \,|$ and $\Phi(\omega)$, not from a semilog plot such as Fig. 5. Estimates, where needed (such as end points of curves), were made on the linear plot.

The impulse response computed by the approximation (6) is shown in Fig. 7.

One notices that the graphical transform yields a nonzero value at $t = 0$, and suggests a nonzero response for $t < 0$. The reason for this might be one of several: (a) the phase and amplitude data of Fig. 5 may not be compatible to satisfy the assumptions made about the system; (b) the data of Fig. 5 suggest that the amplitude response may be band-limited, and it was so treated in the computation; (c) the quantization used in (6) may introduce an error in the calculation of $h(t)$.

Of these three possibilities, the first two seem the more likely sources of discrepancy. The phase data in Fig. 3 suggest that at very low frequencies the phase difference between the displacements of the membrane and stapes is essentially zero. We know, however, that the scalas vestibuli and tympani communicate at the helicotrema. Consequently, a constant displacement of the stapes cannot sustain a constant displacement of the membrane. This argues, therefore, that the amplitude of membrane displacement must go to zero as zero frequency is approached, and the frequency-domain transform of displacement must have at least one zero at the origin of the complex frequency plane. If this is the case, and if the transform is minimum phase, the phase response near zero frequency must be at least $\pi/2$. Intuitively, too, it appears that constant displacement near the helicotrema requires constant velocity of the

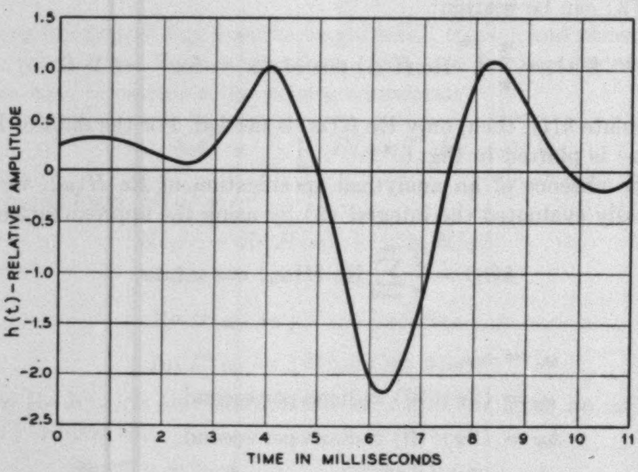


Fig. 7 — Impulse response of the point on the basilar membrane characterized by the amplitude and phase data of Fig. 5. The inverse Fourier transform is obtained by graphical integration of the experimental frequency-domain data.

stapes, arguing again for a derivative relationship between displacements at low frequencies. It seems likely then, that, as low frequencies are approached, the phase of the membrane displacement begins to lead that of the stapes and at zero frequency goes to $\pi/2$. Measurement of the phase relations at low frequencies undoubtedly is difficult, owing to minuscule displacement of the membrane.

In connection with possibility (b), the amplitude data in Fig. 2 suggest that the membrane displacement is essentially band-limited and diminishes to zero for frequencies below about 0.05 and above about 2.0 times the resonant frequency. This should be interpreted, however, with an appreciation of the magnitudes of displacement being observed (on the order of 10^{-4} cm) and the precision attaining thereto. In the graphical transformation, an effort was made to follow the experimental indications as exactly as possible. The amplitude function was treated as mathematically band-limited and was considered to have zero value for frequencies above 400 cps and below 5 cps. This probably is not realistic for the physical system.

Nevertheless, the inverse transform of the experimental data will provide a helpful guide for appraising the responses of the models to be developed in the next section.

IV. MODELS FOR BASILAR MEMBRANE DISPLACEMENT

A model for calculating the displacement of the basilar membrane at a given point must fit the frequency-domain data shown in Figs. 2 and 3. The response curves for various points along the membrane are not unlike those of bandpass filters having relatively sizable in-band delays. The peak values of the curves of Fig. 2 have been normalized to unity, but, as we recall from the previous discussion and from Fig. 4, the peak response rises at about 5 db/octave in the frequency range up to 1000 cps. Above about 2000 cps, the peak response probably falls at something around 12 db/octave, and the stapes displacement is no longer in-phase with the pressure at the drum.

If the data of Figs. 2 and 3 are normalized with respect to the frequency of the maximum response, the curves of Figs. 8 and 9 are obtained, respectively.* Except for the 150-cps case, the phase curves are estimated by reading points vertically from Fig. 3. The 150-cps curve is a single complete phase response published by Bekesy.[3]

---

* I have replotted these data as carefully as possible from the published curves of Bekesy. In reviewing the literature a small discrepancy appears between the amplitude curves published in Akustische Zeitschrift and those which appear later in the *Handbook of Experimental Psychology*. I judged this to be due to rounding and smoothing in redrafting the latter, and hence gave more weight to the earlier data.

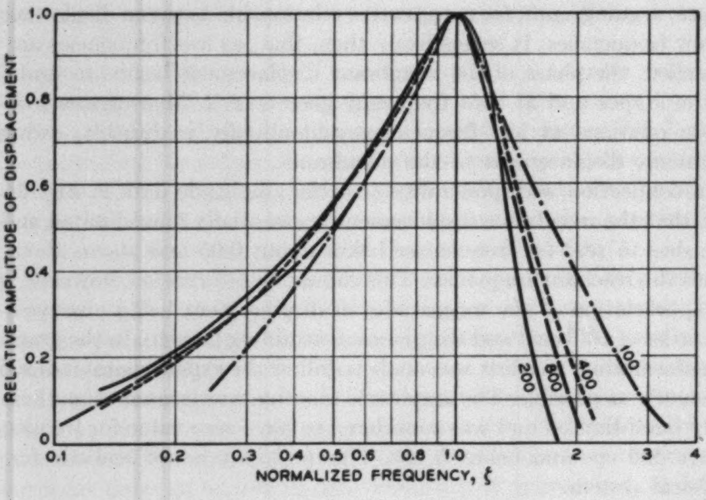


Fig. 8 — The experimental displacement data of Fig. 2 plotted with frequency normalized in respect to the frequency of maximum displacement.

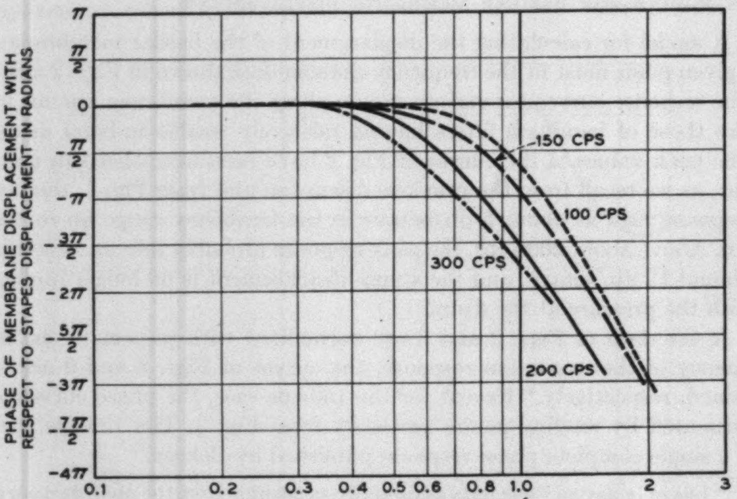


Fig. 9 — Phase responses deduced from data in Fig. 3. Frequency is normalized as in Fig. 8.

One notices that, except for the 100-cps case, the amplitude curves fall close together and represent resonances whose bandwidths are essentially constant percentages of the resonant frequencies (i.e., constant "Q"). The 100-cps curve is slightly broader than the others. The lower skirt of the amplitude curves rises at about 6 db/octave, while the upper skirt falls at approximately 20 to 30 db/octave. The total phase change in passing through a resonance is of the order of 3π. The phase curves for the lower frequency points have the greater slopes (i.e., $d\Phi/d\omega$) inside the passbands, and the delay for the lower frequency points is therefore greater. (This is, of course, as it should be, since the time required to propagate energy from the eardrum to points near the apical end of the membrane is greater than it is for points lying at the basal end.)

As a minor digression, it is interesting to notice that the slopes of the phase curves in the vicinity of resonance indicate delay values about twice as large as the transit times measured by Bekesy.[4] Measuring the slopes of the phase curves in this region (again, from the linear plot) yields:

Resonant Frequency, f	Phase Delay, $d\phi/d\omega$	$2\pi f(d\phi/d\omega)$
100 cps	11.8 msec	7.4 radians
150	7.2	6.8
200	6.4	8.0
300	4.5	8.5

These times represent the delays of the frequency components containing the greatest portion of the stimulus energy, and do not represent the times at which a response first appears (i.e., transit times). Looking back at the graphically determined impulse response for the 200-cps point (Fig. 7), one sees that the greatest displacement occurs at approximately 6.3 milliseconds. The time at which the response essentially begins is of the order of 2.5 milliseconds, which is in close agreement with Bekesy's measurements. It is also interesting to note in passing that the product of resonant frequency and delay near resonance (i.e., the third column) is roughly constant. This fact will be utilized in adjusting the phase response of the models.

To return to the question of fitting a function to the frequency-domain data, at least for the frequency range below 1000 cps, let us consider a model whose Laplace transform is the ratio of rational polynomials. There will be, of course, an infinite number of possibilities for fitting the data, depending upon the criterion and precision of fit. We would, however, like to have an approximation that is both computationally simple and hopefully adequate to explain certain subjective results in pitch-matching. Any criterion of fit must ultimately have its roots in psychoacoustic phenomena. Since such cannot be specified at this time, it would seem that conventional curve-fitting techniques and least-squares criteria might be discarded in favor of a basically intuitive approach.

The skirt slopes of the amplitude curves suggest a frequency function that has a simple zero in the vicinity of the origin of the complex frequency plane, and a denominator whose degree is about four or five greater than that of the numerator. The relationship between the real and imaginary parts of its complex conjugate poles ought to be such as to maintain the constant-percentage bandwidth character of the responses. The amplitude at resonance ought to vary in the manner prescribed earlier, and the phase and delay characteristics presumably should be representative of the experimental data. (The question of phase at low frequencies will necessarily receive some further consideration.)

As one of the simpler possibilities for approximating the amplitude and phase data, consider a function having two pairs of synchronously tuned complex-conjugate poles, one negative-real axis pole, and one negative-real axis zero near the origin. Adorned with necessary constants, such a function has a Laplace transform:

$$F_1(s) = c_1\beta^{4+r}\left(\frac{s+\epsilon}{s+\gamma}\right)\left[\frac{1}{(s+\alpha)^2+\beta^2}\right]^2 e^{-sT}, \qquad (7)$$

where:

c_1 is a positive real scale factor which yields the appropriate absolute value of displacement;

β^{4+r} is a factor that produces the proper variation in amplitude of resonance with resonant frequency (if, as previously suggested, a figure of 5 db/octave rise in the resonant peak is accepted, then $r = 0.83$);

e^{-sT} is a delay factor (T seconds) to bring the phase response into line with the experimental phase data.

The function has second-order poles at $s = -\alpha \pm j\beta$, a simple pole at $s = -\gamma$ and a simple zero at $s = -\epsilon$. By virtue of the constant-percentage bandwidth properties of the membrane resonances, we let β and α be related by a constant: $\beta = k\alpha$. The value of the function for real frequencies (i.e., $s = j\omega$) is:

$$F_1(j\omega) = c_1\beta^{4+r}\left(\frac{\epsilon+j\omega}{\gamma+j\omega}\right)\left[\frac{1}{\left(\beta^2+\frac{\beta^2}{k^2}-\omega^2\right)+j\frac{2\beta}{k}\omega}\right]^2 e^{-j\omega T}. \qquad (8)$$

As with the experimental data, it is convenient to work with frequency normalized. Let $\zeta = (\omega/\beta)$.* Then (8) becomes:

$$F_1(j\zeta) = c_1\beta^r \left(\frac{\frac{\epsilon}{\beta} + j\zeta}{\frac{\gamma}{\beta} + j\zeta}\right)\left[\frac{1}{\left(1 + \frac{1}{k^2} - \zeta^2\right) + j\,\frac{2}{k}\,\zeta}\right]^{-2} e^{-j\zeta\beta T}. \tag{9}$$

One notices that fitting the phase and amplitude data of Bekesy near to zero frequency presents somewhat of a dilemma (as it does with all other minimum-phase functions that we have considered). To diminish the amplitude response at low frequencies, one needs the zero of the function close to the origin. Although the phase at zero frequency obviously remains zero so long as the function zero is in the left-half plane, the phase "bulges" appreciably positive at low frequencies if the zero is placed too close to the origin. By empirical adjustment of the parameters, a compromise position was obtained for the zero, and corresponding values for k, T and γ were deduced. The values arrived at are:

$$\begin{aligned} \frac{\epsilon}{\beta} &= 0.1, & k &= 2.0, \\ \frac{\gamma}{\beta} &= 1.0, & T &= \frac{3\pi}{4\beta}\ \text{seconds.} \end{aligned} \tag{10}$$

In order to match phase responses, one notices that the delay, T, is taken to vary inversely with the resonant frequency, β. For the constant chosen, the added delay at 100 cps, for example, is approximately 4 milliseconds. This delay, in conjunction with the ω-dependent delay, is in reasonable agreement with Bekesy's measurements of transit time down the membrane.

A plot of

$$\frac{|\,F_1(j\zeta)\,|}{|\,F_1(j\zeta_{\max})\,|},$$

where $\zeta_{\max}$ is the frequency of peak displacement, is given in Fig. 10.† The hatched region represents, for comparison, the variability among the

---

* This normalizes real frequency with respect to the imaginary part of the pole frequency. The latter is not necessarily the same as the frequency of maximum response.

† Note that for the present parameters the resonant peak does not fall exactly at $\zeta = 1.0$, but more nearly at $\zeta = 0.95$.

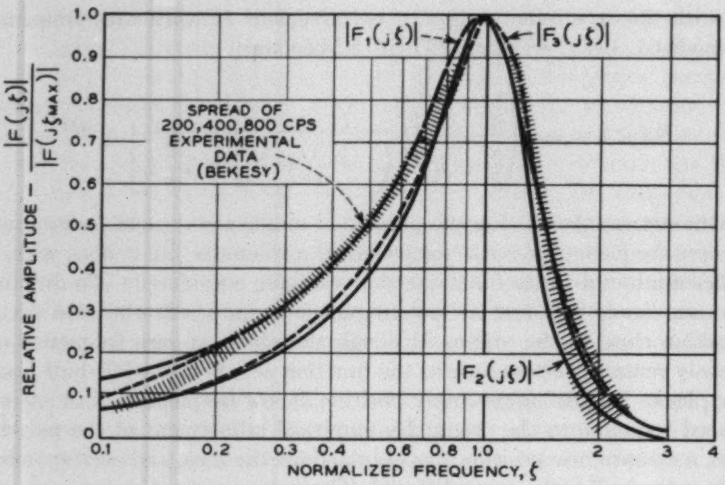


Fig. 10 — Frequency responses of the models compared with experimental data.

200, 400 and 800 cps curves of Fig. 8. A plot of $\angle F_1(j\zeta) = \Phi_1(j\zeta)$ is given in Fig. 11.

If the experimental phase data at low frequencies are not taken too seriously, and the phase of (9) allowed to approach $\pi/2$, then the zero might be placed at the origin (i.e., $\epsilon = 0$). The amplitude response for this situation is shown by the dashed portion of the $|F_1(j\zeta)|$ curve in Fig. 10.

At high frequencies, function (9) attenuates as ζ^{-4}, or at about 24 db/octave. Some of Bekesy's data indicate attenuations slightly greater than this. As another possibility, therefore, a function having a simple zero at the origin and third-order, complex-conjugate poles was considered. Its Laplace transform is:

$$F_2(s) = c_2\beta^{5+r} \frac{s}{[(s+\alpha)^2+\beta^2]^3} e^{-sT}, \tag{11}$$

where the constants are defined in a manner similar to (7). The real frequency response in terms of normalized frequency is:

$$F_2(j\zeta) = c_2\beta^r \frac{j\zeta}{\left[\left(1+\frac{1}{k^2}-\zeta^2\right)+j\frac{2}{k}\zeta\right]^3} e^{-j\zeta\beta T}. \tag{12}$$

A reasonable fit to the resonant bandwidth is obtained for $k = 2.0$ with

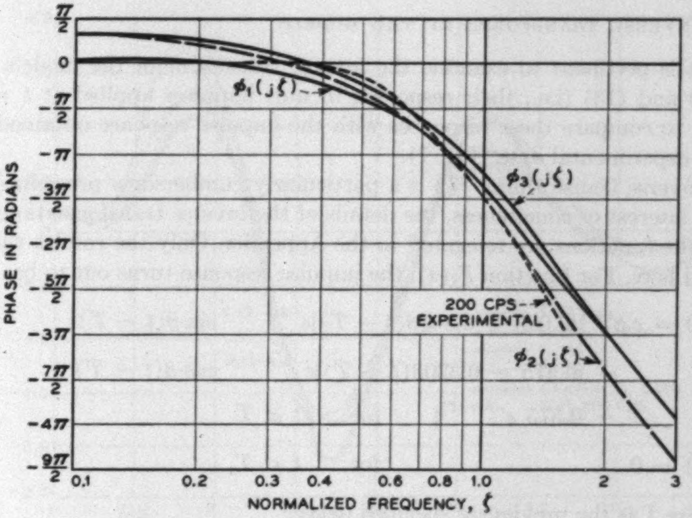


Fig. 11 — Phase responses of the models.

$\beta T = 3\pi/4$, as before. For these values, a plot of $|F_2(j\zeta)|/|F_2(j\zeta_{\max})|$ is given in Fig. 10 and $\angle F_2(j\zeta)$ is given in Fig. 11.

With a thought toward inverse transformations for the approximating functions, one function that provides a respectable fit and has a particularly simple inverse transform is the following:

$$F_3(s) = c_3\beta^{4+r}\,\frac{s^2 + 2\alpha s + \left(\alpha^2 - \dfrac{\beta^2}{3}\right)}{[(s+\alpha)^2 + \beta^2]^3}\,e^{-sT}. \tag{13}$$

Or, in terms of the normalized real frequency,

$$F_3(j\zeta) = c_3\beta^r\,\frac{\left(\dfrac{1}{k^2} - \dfrac{1}{3} - \zeta^2\right) + j\,\dfrac{2}{k}\,\zeta}{\left[\left(\dfrac{1}{k^2} + 1 - \zeta^2\right) + j\,\dfrac{2}{k}\,\zeta\right]^3}\,e^{-j\zeta\beta T}. \tag{14}$$

This function has simple zeros at $s = \alpha(-1 \pm k/\sqrt{3})$ and third-order poles at $s = \alpha(-1 \pm jk)$. The function obviously becomes nonminimum phase for $k > \sqrt{3}$. Because the separation between zeros is $2k/\sqrt{3}$, the zero at $s = \alpha(-1 + k/\sqrt{3})$ has the greatest influence on amplitude response for the minimum phase conditions (i.e., $k \lesseqgtr \sqrt{3}$). For values of $k = 1.7$ and $\beta T = 3\pi/4$, the amplitude and phase responses of (14) are shown in Figs. 10 and 11, respectively.

V. INVERSE TRANSFORMS OF THE MODELS

It is pertinent to examine the inverse transforms of the models (7), (11) and (13) (i.e., their responses to unit impulses applied at $t = 0$) and to compare these responses with the impulse response obtained for the experimental data (Fig. 7).

Inverse transforming (7) is a particularly cumbersome procedure. In the interest of conciseness, the details of the inverse transformations for all the functions are relegated to the Appendix. Only the results will be used here. For function $F_1(s)$, the impulse response turns out to be:

$$\begin{aligned} f_1(t) &= c_1\beta^{1+r}\{[0.033 + 0.360\beta(t - T)]e^{-\beta(t-T)/2} \sin \beta(t - T) \\ &\qquad + [0.575 - 0.320\beta(t - T)]e^{-\beta(t-T)/2} \cos \beta(t - T) \\ &\qquad - 0.575\, e^{-\beta(t-T)}\} \qquad \text{for} \quad t \geqq T \\ f_1(t) &= 0 \qquad \text{for} \quad t < T, \end{aligned} \tag{15}$$

where T is the previously specified delay.

In a similar manner, the inverse transform of $F_2(s)$ is:

$$\begin{aligned} f_2(t) &= \\ &\frac{c_2\beta^{1+r}}{8}\left[\left\{\frac{[\beta(t - T)]^2}{2} + \beta(t - T) - \tfrac{3}{2}\right\} e^{-\beta(t-T)/2} \sin \beta(t - T)\right. \\ &\qquad \left. + \{-[\beta(t - T)]^2 + \tfrac{3}{2}\beta(t - T)\}e^{-\beta(t-T)/2} \cos \beta(t - T)\right] \\ &\qquad\qquad \text{for} \quad t \geqq T, \\ f_2(t) &= 0 \qquad \text{for} \quad t < T. \end{aligned} \tag{16}$$

As indicated earlier, the inverse transform of $F_3(s)$ is particularly simple, this being the principal reason for presenting its fit. Its inverse is:

$$\begin{aligned} f_3(t) &= \frac{c_3\beta^{1+r}}{6}[\beta(t - T)]^2 e^{-\beta(t-T)/1.7} \sin \beta(t - T) \qquad \text{for} \quad t \geqq T \\ f_3(t) &= 0 \qquad \text{for} \quad t < T. \end{aligned} \tag{17}$$

For comparison purposes, the impulse responses $f_1(t)$, $f_2(t)$ and $f_3(t)$ are plotted in Fig. 12, together with the graphically determined response of Fig. 7. In this plot relative delays have been equalized to compare waveforms. Because the scale constants c_1, c_2 and c_3 have not been taken into account, the amplitude scales for the different curves are relative. The curves have been plotted, however, for approximately equal

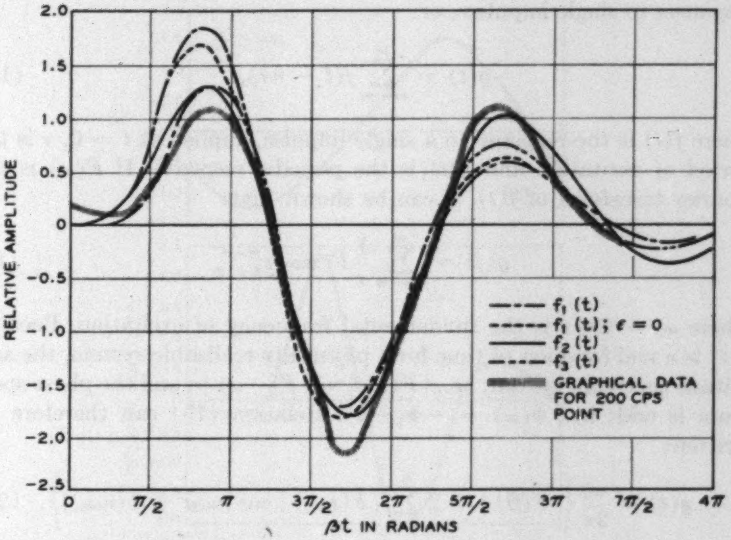


Fig. 12 — Impulse responses of the models. These displacement functions are the inverse transforms of the frequency-domain data in Figs. 10 and 11. Time delay has been equalized to compare waveforms. Locations of absolute origins are given in the text.

peak-to-peak values. The fits to the experimental data do not seem unrealistic, in view of the questions raised earlier. One notices that, in most instances, the positive impulses produce the greatest deflection in the negative direction. Equalization of the delays to bring the curves into coincidence were such as to make the absolute origins ($\beta t = 0$) for each response the following number of radians to the left:

Function	Radians to Absolute Origins
200 cps, experimental	2.3
$f_1(t)$	1.9
$f_2(t)$	2.4
$f_3(t)$	1.5

Of the functions displayed, $f_2(t)$ and $f_3(t)$ appear to fit the graphically derived impulse response better than $f_1(t)$ does. In the frequency domain, however, $F_1(s)$ appears to afford the slightly better fit.

VI. RESPONSE OF MODELS TO PERIODIC IMPULSE EXCITATION

If an excitation of periodic unit impulses is delivered to a linear system, the periodic response is a doubly infinite, linear superposition of responses to single impulses, or:

$$g(t) = \sum_{n=-\infty}^{\infty} f(t - n\tau), \tag{18}$$

where $f(t)$ is the response to a single impulse, applied at $t = 0$, τ is the period of excitation and $g(t)$ is the periodic response. If $F(\omega)$ is the Fourier transform of $f(t)$, it can be shown that:

$$g(t) = \sum_{n=-\infty}^{\infty} \frac{1}{\tau} F(n\omega_0)e^{jn\omega_0 t}, \tag{19}$$

where $\omega_0 = 2\pi/\tau$ is the fundamental frequency of excitation. Because $g(t)$ is a real function of time for a physically realizable system, the amplitude spectrum is even; i.e., $| F(\omega) | = | F(-\omega) |$; and the phase spectrum is odd; i.e., $\Phi(\omega) = -\Phi(-\omega)$. Relation (19) can therefore be written:

$$g(t) = \frac{\omega_0}{2\pi}\left\{| F(0) | + 2\sum_{n=1}^{\infty} | F(n\omega_0) | \cos [n\omega_0 t + \Phi(n\omega_0)]\right\}. \tag{20}$$

By way of example, let us look at the response of function $F_1(\omega)$ [see (8)] to an excitation of periodic impulses. Suppose we first take the case where $F_1(\omega)$ specifies a point on the membrane tuned to the fundamental frequency of excitation. Let the resonant frequency of the point be $\beta_x = \omega_0$. Then $\zeta = \omega/\omega_0 = n\omega_0/\omega_0 = n$ and $F_1(n\omega_0) = F_1(\zeta = n)$, and the periodic response is:

$$g_x(t) = \frac{\beta_x}{2\pi}\left\{F_1(\zeta = 0) + 2\sum_{n=1}^{\infty} | F_1(\zeta = n) | \cos [n\beta_x t + \Phi_1(\zeta = n)]\right\}. \tag{21}$$

As determined in previous calculations, values of $F_1(\zeta)$ are:

n	ζ	$\left\lvert\frac{F_1(\zeta)}{c_1\beta_x^{\tau}}\right\rvert$	$\phi(\zeta)$, degrees
0	0	0.06	0
1	1	0.67	−248
2	2	0.08	−534
3	3	0.01	−706

Obviously, in this case the displacement response of the membrane is principally fundamental, the second harmonic being slightly more than one-tenth the amplitude of the fundamental. A plot, on a relative amplitude scale, of these first four terms is shown in Fig. 13(a).

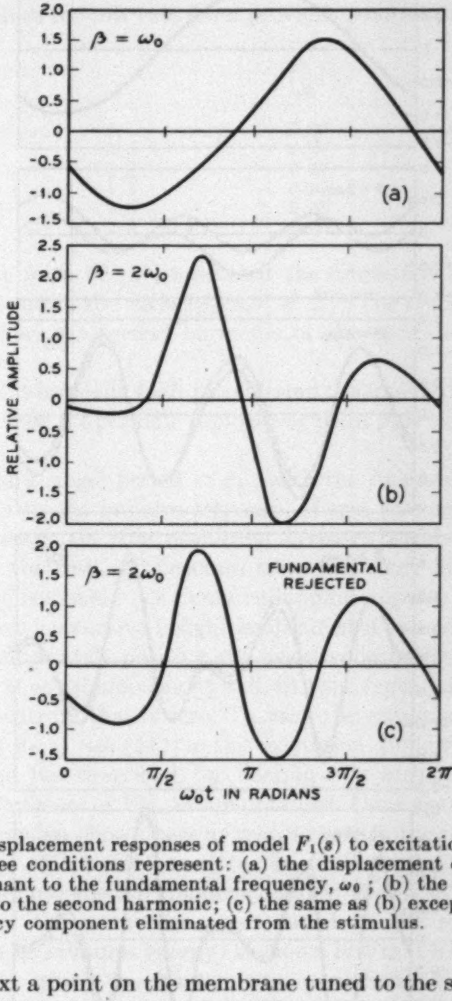


Fig. 13 — Displacement responses of model $F_1(s)$ to excitation by periodic impulses. The three conditions represent: (a) the displacement of a point on the membrane resonant to the fundamental frequency, ω_0; (b) the displacement of a point resonant to the second harmonic; (c) the same as (b) except with the fundamental frequency component eliminated from the stimulus.

Consider next a point on the membrane tuned to the second harmonic of the stimulus (i.e., $\beta_y = 2\omega_0 = 2\beta_x$). Then $\zeta = \omega/2\omega_0 = n\omega_0/2\omega_0 = n/2$ and $F_1(n\omega_0) = F_1(\zeta = n/2)$. In this case:

$$g_y(t) =$$

$$\frac{\beta_y}{4\pi}\left\{F_1(\zeta = 0) + 2\sum_{n=1}^{\infty}\left|F_1\left(\zeta = \frac{n}{2}\right)\right|\cos\left[n\frac{\beta_y}{2}t + \Phi_1\left(\zeta = \frac{n}{2}\right)\right]\right\}. \tag{22}$$

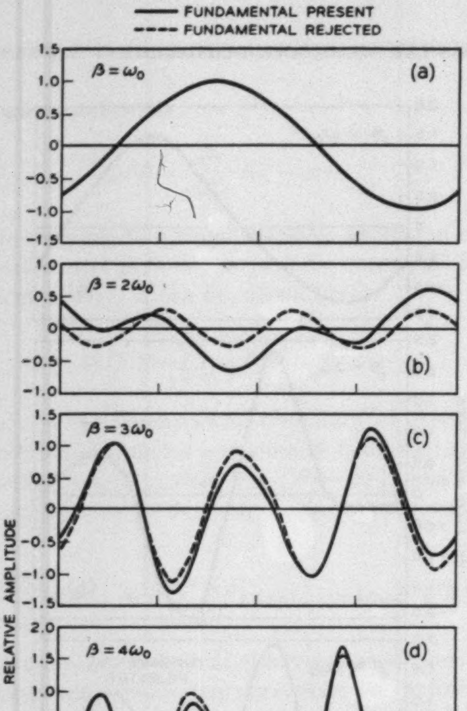


Fig. 14 — Displacement responses of model $F_2(s)$ to periodic excitation by alternate positive and negative impulses. The five conditions represent the displacements of membrane points respectively resonant to: (a) fundamental frequency; (b) second harmonic; (c) third harmonic; (d) fourth harmonic and (e) fifth harmonic. The dashed curves are the displacements when the fundamental component is eliminated from the stimulus.

Functional values for this case from previous computations are:

n	ζ	$\left\lvert\frac{F_1(\zeta)}{c_1\beta_y{}^r}\right\rvert$	$\phi(\zeta)$, degrees
0	0	0.06	0
1	0.5	0.37	−69
2	1.0	0.67	−248
3	1.5	0.27	−422
4	2.0	0.08	−534
5	2.5	0.03	−626

Because of the form of (9), note that the amplitude scale factors for $g_y(t)$ and $g_z(t)$ are in the ratio $(\beta_y/\beta_z)^r = 2^r$.* The response $g_y(t)$ of the point resonant at the second harmonic of the excitation is plotted in Fig. 13(b).

If the stimulus is ideally high-pass filtered to remove the dc and fundamental terms, then the periodic response at point β_y is that shown in Fig. 13(c).

The shape of a single period at β_y, with the fundamental present, is already similar to the impulse response. If one examines points tuned higher in frequency, the time resolution increases because the bandwidth increases, and the response becomes more and more identifiable as repeated impulse responses (i.e., nonoverlapping impulse responses).

An even more instructive insight is obtained if one considers periodic excitation by alternately positive and negative impulses. Such a train is odd-harmonic in equal amplitude† and, like the repeated positive pulses, has a phase spectrum that is zero. To vary the example, let us consider the response of $F_2(s)$ [see (11)] to this excitation. Following an approach identical to that just described, but dealing only with odd spectral components, the responses of Fig. 14 are obtained. Once again we recall that the amplitude scales, shown here as relative, are in the ratio β^r.

The response of a point tuned to the fundamental is essentially a sinusoid at the fundamental frequency and is shown in Fig. 14(a). The displacement of the membrane point tuned to the second harmonic (where there is no stimulus energy) is shown in Fig. 14(b). It exhibits a displacement in which the fundamental periodicity can be discerned when the fundamental component is present. Without the fundamental the response is relatively low-amplitude third harmonic. The point tuned to the third harmonic, Fig. 14(c), displaces essentially at the third harmonic frequency whether the fundamental is present or absent. The point tuned to the fourth harmonic, Fig. 14(d), begins to exhibit fundamental periodicity again, regardless of whether fundamental is present or not. The point tuned to the fifth harmonic, Fig. 14(e), yields a response which is very nearly nonoverlapping, superposed impulse responses.

---

* The implication here, of course, is that we are still dealing with frequency ranges below 1000 cps, where the membrane resonances are assumed to vary in peak displacement, as previously discussed.

† The equal-amplitude spectral lines have twice the amplitude of those for repeated positive impulses.

Quantification of the membrane displacement in this manner offers a basis for a number of useful speculations on the perception of periodic pulses.

VII. CONCERNING RELATIVE AMPLITUDES OF DISPLACEMENT

Since relative amplitude of displacement may be of importance in the conversion of membrane displacement into nervous activity, it is worthwhile to examine amplitude relations further. We have seen that, if the membrane is excited with periodic impulses at a fundamental frequency to which a point near the apical (low-frequency) end is resonant, this point executes a displacement which is nearly the fundamental sinusoid. A point toward the basal (high-frequency) end, whose resonance curve embraces a substantial number of harmonics, yields a periodic response, which is essentially a succession of negligibly-overlapping impulse responses. Because such points respond simultaneously (except for transit delay), and because their peak amplitudes have implications in hypotheses about pitch perception, let us compare the peak amplitudes of a "fundamental-responding" point with that of an "impulse-responding" point. For the sake of varying the examples further, let us work with model $F_3(s)$, in (13), and its impulse response $f_3(t)$, in (17). We are interested in the absolute extremum of (17). The times of the extrema can be found by differentiating (17), setting to zero and solving, which gives:

$$t_{\max} = \frac{1}{\beta} \tan^{-1}\left[\frac{1.7\beta(t-T)}{\beta(t-T)-3.4}\right], \qquad t > T. \tag{23}$$

The envelope maximum occurs at:

$$t_{\max\ \mathrm{envel}} = \left(\frac{3.4}{\beta} + T\right). \tag{24}$$

It is not necessary to solve the trancendental relation (23), since we already have (17) plotted to a reasonable precision in Fig. 12. Using the latter data, we get for the absolute maximum of $f_3(t)$,

$$|f_3(t)|_{\max} = \frac{c_3\beta_p^{\,1+r}}{6}(1.4) = (0.23)c_3\beta_p^{\,1+r}, \tag{25}$$

where the subscript p denotes a point toward the high-frequency end of the membrane. In a parallel manner, the amplitude of a point, q, tuned to the fundamental frequency can be obtained from relation (20). In this case, $\beta_q = \omega_0$ and

$$\begin{aligned} | g_3(t) |_{\text{fund}} &\cong \frac{\omega_0}{2\pi} 2 \, | F_3(\zeta = 1) | \\ &\cong \frac{\omega_0}{\pi} c_3 \beta_q^{\,r} (0.83) \\ &\cong c_3 \beta_q^{\,1+r} (0.26). \end{aligned} \tag{26}$$

The ratio of these two peak displacements is, therefore

$$R_3 = \frac{| f_3(t) |_{\max}}{| g_3(t) |_{\text{fund}}} = (0.88) \left(\frac{\beta_p}{\beta_q}\right)^{1+r}. \tag{27}$$

If the same computations are made for the other two models, $F_1(s)$ and $F_2(s)$, the ratios are:

$$\begin{aligned} R_1 &= (0.80) \left(\frac{\beta_p}{\beta_q}\right)^{1+r}, \\ R_2 &= (0.82) \left(\frac{\beta_p}{\beta_q}\right)^{1+r}. \end{aligned} \tag{28}$$

Since $\beta_p > \beta_q$ and since the experimentally determined exponent $r \approx 0.8$, the peak amplitudes of the impulse-responding points exceed those of the fundamental-responding points, at least in the frequency range below 1000 cps (i.e., roughly over the apical half of the membrane).

VIII. EVALUATION OF SCALE CONSTANTS c_1, c_2 AND c_3

Bekesy's data show that the maximum deflection of the basilar membrane at a frequency of 1000 cps and a sound pressure level of 134 db referred to 0.0002 dyne/cm² (i.e., 10^3 dynes/cm²) is approximately 10^{-4} cm. His measurements also indicate that the mechanical functioning of the middle and inner ear is essentially linear to the threshold of feeling. In the models, therefore, the constants c_1, c_2 and c_3 should be chosen to provide peak displacements at resonance equal to

$$(10^{-7} \text{ cm}^3/\text{dyne}) \left[\frac{\beta}{2\pi(1000)}\right]^r.$$

The amplitude responses of the models at resonance are:

$$\begin{aligned} \mid F_1(\zeta = 1.0) \mid &= c_1\beta^r(0.66), \\ \mid F_2(\zeta = 1.0) \mid &= c_2\beta^r(0.92), \qquad (29) \\ \mid F_3(\zeta = 1.0) \mid &= c_3\beta^r(0.83). \end{aligned}$$

The values of the constants, therefore, should be:

$$\begin{aligned} c_1 &= \frac{10^{-7}}{(0.66)[2\pi(1000)]^r}, \\ c_2 &= \frac{10^{-7}}{(0.92)[2\pi(1000)]^r}, \qquad (30) \\ c_3 &= \frac{10^{-7}}{(0.83)[2\pi(1000)]^r}. \end{aligned}$$

IX. APPLICATION TO PITCH PERCEPTION

As suggested at the outset, the present computations were precipitated by a particular need. In drafting a paper to report two earlier experiments on pitch perception,[1,2] it became painfully obvious, as soon as the discussion section was reached, that little quantitative basis existed for interpreting the subjective data. The models described here were developed in an attempt to alleviate this situation.

In the pitch experiments it became necessary to explain how three different modes of pitch perception arise when periodic pulse trains stimulate the ear. One mode ascribes a pitch to the stimulus equal to the pulse rate, regardless of the polarity pattern of the train; in other words, positive pulses (condensations) are not discriminated from negative pulses (rarefactions). A second mode ascribes a pitch equal to the mathematical fundamental whether energy is present at this frequency; this mode includes the situation which has been labeled "residue" phenomenon. The third mode assigns a pitch equal to the frequency of the lowest spectral component present in the stimulus.

The first mode characteristically operates at low values of pulse rate (usually below 100 pps in unmasked situations). The second usually obtains for fundamental frequencies in the approximate range 200 to 500 cps. The third seems to hold for fundamental frequencies around 1000 cps and higher when the lowest-frequency component is rejected by HP filtering.

Without launching into the details of the psychophysical experiments, the applicability of the models to the perception of pulses can at least be indicated. It is of consequence, for example, to ascertain to what extent the subjective pitch modes are manifested in the mechanical operation of the cochlea. Looking again at Fig. 14, one can observe displacement patterns that might be considered favorable for giving rise to the pitch modes just outlined. This presumes, of course, certain hypotheses about the mechanism of converting displacement information into electrical discharges in the nerve fiber. A discussion of these important details, however, is more appropriate in another place. Even so, Fig. 14 suggests several things.

When the membrane is excited over most of its length by a periodic pulse stimulus, the higher-frequency portion probably is effective in supplying only pulse-rate information, no matter what the polarity pattern of the train. In this region of the membrane the pulses are well resolved in time (i.e., the displacement is essentially nonoverlapping impulse responses), and the "overshoot" of the response to each pulse is substantial. Under certain assumptions about the transduction of displacement into nervous activity, the latter fact can be construed as favorable for eliciting nerve volleys in synchrony with each pulse.*

Information on fundamental frequency might be manifested in two ways: (a) If the fundamental component is present in the stimulus, then the point on the membrane tuned to the fundamental responds strongly with near sinusoidal displacement. (b) If, on the other hand, the fundamental is absent, the lowest-frequency part of the membrane receiving excitation will embrace a small number of spectral lines within its frequency response. Its displacement generally will exhibit the fundamental periodicity in a form favorable for triggering one nerve volley per fundamental period.

So far these comments have not considered the importance of relative amplitudes of displacement. This question appears to be of particular consequence in evoking the second, or fundamental, pitch mode. Although the indications are that most significant neural information originates from the point of greatest displacement, there is evidence that subjects may give preference to the fundamental mode over the pulse-rate mode even though the former may be correlated with smaller membrane displacements than is the latter. Relative amplitudes of displacement very likely undergo nonsimple transformations in the neural conversion process.

Still open, too, is the question of the third pitch mode. Although our models are limited to the frequency range below 1000 cps (because they do not adequately account for middle-ear transmission above this frequency), an explanation, fabricated of flimsy substance, can be suggested for the third mode. Bekesy's data suggest that the amplitude of maximal displacement of the membrane falls appreciably (about 12 db/octave or more) for frequencies above 1000 cps. In this region, then, that part of the membrane responding to the lowest-frequency component would exceed in amplitude those parts responding to higher-frequency components. If amplitude of displacement is at all important in the conversion process (and it most probably is), then the third mode is favored *provided* the lowest-frequency component is not too high in harmonic number. As indicated earlier, the third mode has been observed when either the fundamental, or the fundamental and second harmonic, is rejected from the stimulus. This mode has obtained in our pitch-matching experiments for fundamentals in the frequency range around 1000 cps and slightly higher.

---

* There also is evidence that the transduction may be sensitive to spatial derivatives of displacement as well as to displacement. This, too, could facilitate perception of the pulse-rate mode.

One final comment is of interest along these same lines. It has been reported in the literature that if a periodic train of positive pulses is high-pass filtered at around 3000 and 4000 cps, one hears a "residue" pitch equal to the fundamental frequency. Our models suggest, however, a response more nearly correlated with pulse rate. If one uses a stimulus in which pulse rate and fundamental frequency are confounded (as with positive pulses), then the former result might obtain. If, on the other hand, a stimulus such as alternate positive and negative pulses were used, the subjective impression may well be that of pulse rate. If the latter is in fact the case, then a fundamental "residue" pitch does not exist for this condition.*

X. ACKNOWLEDGMENT

I wish to thank J. L. Perry of the Acoustics Research Department of Bell Telephone Laboratories for his capable assistance in performing many of the numerical evaluations and in plotting the results.

APPENDIX

Inverse Transforms for $F_1(s)$, $F_2(s)$ and $F_3(s)$

When the function $F_1(s)$ of (7) is disencumbered of its constants, the problem of inverse transformation amounts to calculating the inverse transform of:

---

* Since drafting this paper, I have set up the latter experiment and listened to alternate positive and negative pulses HP-filtered at 3000 and 4000 cps. I made pitch matches fairly consistently at the pulse rate. A second listener, on the other hand, made matches that were generally higher than the pulse rate, suggesting that my preconceived notions may have influenced my data. It is unequivocal, however, that one would not match to the fundamental frequency.

$$\begin{aligned} K_1(s) &= \left(\frac{s+\epsilon}{s+\gamma}\right)\left[\frac{1}{(s+\alpha)^2+\beta^2}\right]^2 \\ &= \left[\frac{1}{(s+\alpha)^2+\beta^2}\right]^2 + \left(\frac{\epsilon-\gamma}{s+\gamma}\right)\left[\frac{1}{(s+\alpha)^2+\beta^2}\right]^2 \qquad (31) \\ &= K_a(s) + K_b(s). \end{aligned}$$

The inverses of $K_a(s)$ and $K_b(s)$ can be obtained in the usual manner by making partial fraction expansions in terms of the singularities, account being taken of the order of the poles, and evaluating the residues in each pole. Or, having got the inverse for $K_a(s)$, the inverse for $K_b(s)$ can be computed from:

$$K_b(t) = [(\epsilon-\gamma)e^{-\gamma t}] * [\mathcal{L}^{-1}K_a(s)], \qquad (32)$$

where $*$ indicates convolution.

For the present case these standard procedures prove rather cumbersome and messy. Because of the favorable initial values of the function and its first two derivatives [namely, $k_1(0_+) = k_1'(0_+) = k_1''(0_+) = 0$], derivative relationships can be used to obviate evaluating residues and performing the convolution.* The derivative relations of use here are the following: If the function $f(t)$ has the Laplace transform $F(s)$, then

$$(-1)^n \frac{d^nF(s)}{ds^n} = t^nf(t), \qquad (33)$$

and

$$\frac{d^nf(t)}{dt^n} = s^nF(s) - s^{n-1}f(0_+) - s^{n-2}f'(0_+) - \cdots - f^{n-1}(0_+). \qquad (34)$$

We start with two well-known transform pairs:

$$\frac{1}{(s+\alpha)^2+\beta^2} \rightarrow \frac{1}{\beta}e^{-\alpha t}\sin\beta t = h_1(t), \qquad (35)$$

and

$$\frac{(s+\alpha)}{[(s+\alpha)^2+\beta^2]} \rightarrow e^{-\alpha t}\cos\beta t = h_2(t). \qquad (36)$$

Applying (33) through (36) gives

$$\frac{[(s+\alpha)^2-\beta^2]}{[(s+\alpha)^2+\beta^2]^2} \rightarrow te^{-\alpha t}\cos\beta t = th_2(t). \qquad (37)$$

---

* I am indebted to B. F. Logan of the Acoustics Research Department of Bell Telephone Laboratories, who pointed out to me the utility of the derivative relationships in obtaining transforms for these functions.

One notices that $K_a(s)$ can be expressed as a simple combination of (35) and (37), namely,

$$\frac{1}{[(s+\alpha)^2+\beta^2]^2} = \frac{1}{2\beta^2}\left\{\frac{1}{(s+\alpha)^2+\beta^2} - \frac{(s+\alpha)^2-\beta^2}{[(s+\alpha)^2+\beta^2]^2}\right\} \tag{38}$$

and

$$\frac{1}{[(s+\alpha)^2+\beta^2]^2} \rightarrow \frac{1}{2\beta^2}\,(h_1 - th_2). \tag{39}$$

Application of (34) through (39) gives

$$\frac{s}{[(s+\alpha)^2+\beta^2]^2} \rightarrow h_1\left(\frac{t}{2} - \frac{\alpha}{2\beta^2}\right) + h_2\,\frac{\alpha t}{2\beta^2} \tag{40}$$

and

$$\frac{s^2}{[(s+\alpha)^2+\beta^2]^2} \rightarrow h_1\left(\frac{\alpha^2+\beta^2}{2\beta^2} - \alpha t\right) + h_2\left(\frac{\beta^2-\alpha^2}{2\beta^2}\,t\right). \tag{41}$$

The inverse of $K_a(s)$ is, therefore, (39). The inverse of $K_b(s)$ can be obtained from a partial fraction expansion followed by application of (39), (40) and (41). Expand $K_b(s)$ as:

$$\left(\frac{\epsilon-\gamma}{s+\gamma}\right)\frac{1}{[(s+\alpha)^2+\beta^2]^2} = \frac{A}{(s+\gamma)} + \frac{G(s)}{[(s+\alpha)^2+\beta^2]^2}, \tag{42}$$

where A is a constant and

$$G(s) = (a_0 + a_1 s + a_2 s^2 + a_3 s^3).$$

If A and $G(s)$ are evaluated, one gets

$$\begin{aligned} A &= \left.\frac{(\epsilon-\gamma)}{[(s+\alpha)^2+\beta^2]^2}\right|_{s=-\gamma} = \frac{(\epsilon-\gamma)}{[\gamma^2-2\alpha\gamma+\alpha^2+\beta^2]^2}, \\ a_0 &= \frac{1}{\gamma}\,[\epsilon-\gamma-A(\alpha^2+\beta^2)^2], \\ a_1 &= A[\gamma(4\alpha-\gamma)-2(3\alpha^2+\beta^2)], \\ a_2 &= -A(4\alpha-\gamma), \\ a_3 &= -A. \end{aligned} \tag{43}$$

The inverse transform of $K_b(s)$, therefore, is a summation of terms (39) through (41), with the appropriate multiplicative constants.

Two differentiations (with respect to s) of (35) give the transform pair:

$$\frac{[(s+\alpha)^2-\beta^2/3]}{[(s+\alpha)^2+\beta^2]^3} \rightarrow \frac{1}{6}t^2h_1, \tag{44}$$

which is the function used as the model $F_3(s)$ of (13).

In an essentially parallel manner, one obtains the pair:

$$\frac{s}{[(s+\alpha)^2+\beta^2]^3} \rightarrow \frac{1}{8\beta^4}[h_1(\alpha\beta^2t^2+\beta^2t-3\alpha)+h_2(3\alpha t-\beta^2t^2)]. \tag{45}$$

This is the function used as the model $F_2(s)$ of (11).

REFERENCES

1. Flanagan, J. L. and Guttman, N., Pitch of Periodic Pulses, J. Acoust. Soc. Am., **31**, 1959, p. 123.
2. Flanagan, J. L. and Guttman, N., Pitch of Periodic Pulses Without Fundamental Component, J. Acoust. Soc. Am., **31**, 1959, p. 836.
3. von Bekesy, G., Variations of Phase Along the Basilar Membrane with Sinusoidal Vibrations, J. Acoust. Soc. Am., **19**, 1947, p. 452.
4. von Bekesy, G., Über die Resonanzkurve und die Abklingzeit der verscheidenen Stellen der Schneckentrennwand, Akust. Zeit., **8**, 1943, p. 66; J. Acoust. Soc. Am., **21**, 1949, p. 245.
5. von Bekesy, G., Über die Schwingungen der Schneckentrennwand beim Präparat und Ohrenmodell, Akust. Zeit., **7**, 1942, p. 173; J. Acoust. Soc. Am., **21**, 1949, p. 233.
6. Zwislocki, J., Some Impedance Measurements on Normal and Pathological Ears, J. Acoust. Soc. Am., **29**, 1957, p. 1312.

Erratum

On page 747 of "The Theory and Design of Chirp Radars" in the July 1960 Bell System Technical Journal, the analytical work attributed to A. W. Schelling should correctly be credited to J. C. Schelleng.

Design and Performance of Ultraprecise 2.5-mc Quartz Crystal Units

By A. W. WARNER

(Manuscript received March 29, 1960)

A 2.5-mc crystal unit has been developed for use in a new, extremely stable frequency standard oscillator. A well-balanced design was achieved by using a 30-mm-diameter, plano-convex, polished quartz plate, coated with gold and operated on its fifth overtone. The quartz plate is mounted on its quiescent edge in an evacuated bulb, and achieves a Q of five to six million, representative of the Q of the quartz itself. The temperature coefficient, current coefficient, frequency adjustment tolerance and frequency aging of the crystal unit are all consistent with a frequency stability in the order of one part in 10^{10}. It was necessary to develop polishing methods that would not disturb the crystal structure of the quartz plate and new methods of orienting the crystallographic axes to achieve better temperature coefficient control. New methods of mounting the quartz plate were found that avoid strain and reduce the effects of shock and vibration. The new crystal unit makes possible oscillators characterized by excellent frequency stability, small and uniform aging and straightforward design. For periods up to one month, the frequency stability of such standards compares favorably with that of atomic frequency standards.

I. INTRODUCTION

The quality of a quartz crystal frequency standard is determined by the crystal-controlled oscillator, and particularly by the mechanically vibrating, piezoelectrically excited quartz plate. Special quartz crystal resonators, characterized by high Q, excellent frequency stability under shock and vibration, and small change with time, have been developed for use in a new general-purpose, extremely stable frequency standard.

The development of improved oven and oscillator circuits has contributed substantially to this improved standard, and will be reported in a separate article. Over-all performance of an experimental oscillator has been reported briefly,[1] and similar oscillators are in operation at the National Bureau of Standards (see Section 4.5 of this paper), the Naval Research Laboratory and Bell Telephone Laboratories.

In this article particular emphasis will be given to the quartz resonator, considering (a) the design principles, (b) the development of the present design and the related processing techniques and (c) the properties of the quartz resonator as a circuit element, including its thermal, mechanical and temporal characteristics. In order to limit the scope, a cursory knowledge[2,3] of crystal unit fabrication will be assumed, and only a brief recapitulation of facts already published will be given.

The development of highly stable crystal resonators is a continuing work, because each new achievement in frequency accuracy and stability generates the need for still greater accuracy and stability; to meet these needs, the underlying causes of frequency aging and many other special aspects of the behavior of crystal-controlled oscillators must be more fully understood. A solution of these problems will require further fundamental investigation into the nature of the materials involved.

Such development work for improving quartz oscillators is not likely to be made superfluous in the immediate future by atomic and molecular frequency standards. Atomic standards, whose frequency stability is better than a part per billion for very long periods of time, employ quartz oscillators as part of their circuitry. Thus, their short-time stability is that of the crystal-controlled oscillator. As atomic standards are improved, the need for higher-Q crystal resonators will be increased. Furthermore, as the long-time frequency stability of quartz oscillators is improved, they can be operated for longer periods of time independent of an atomic frequency reference. Use of the oscillator alone would, of course, reduce the size, weight and complexity of the frequency standard.

II. DESIGN PRINCIPLES

In this section the significant parameters in the design of a crystal unit of the highest practical precision are considered, including (a) use of edge-mounted crystal plates operating in high-frequency thickness shear, (b) desirable crystal unit characteristics and their correlation to a well balanced design, (c) the role of quartz plate size and (d) the independence of Q and inductance, and the best choice for the value of the inductance.

There are two basic design concepts in use today for the construction of high-precision quartz crystal units. One makes use of low-frequency, large-size quartz plates supported at nodal points by arrangements of cords and springs or rods, or by soldered wires. Its advantage lies in a high potential Q and a large frequency-determining dimension. Its disadvantage lies in the fact that the mounting structure is part of the frequency-determining, mechanically vibrating portion of the crystal unit, making it unstable with respect to shock and vibration and contributing to frequency aging. The other design concept, and the one to be described here, is the use of edge-mounted crystal units in high-frequency thickness shear operation in order to decouple the mechanically vibrating portion of the quartz plate from the mounting. By use of convex shaping, the mechanical vibration can be confined to the center of the crystal plate, leaving the edge quiescent. Such units can be more closely adjusted to frequency and have improved frequency stability characteristics and other operating advantages as shown in Section IV. The construction details are quite different, relying on carefully designed machines rather than on individual craftmanship.

The crystal unit or resonator for a primary frequency standard must be characterized by high Q, low temperature coefficient of frequency at the operating temperature, low frequency drift with time, low current coefficient, relatively high impedance and small frequency-adjustment tolerance. There is no particular order of importance among these factors, since neglect of any one of them will largely nullify the precision that would be attainable through the use of extreme care with the others.

These requirements, along with performance factors for the oven and circuit, fall naturally into several groupings, with each factor in a group being interrelated with other factors in that group. One combination is the Q of the crystal unit, its frequency accuracy (since frequency adjustment by circuit means is limited by the probable stability of the controlling circuit element) and the phase stability of the oscillator circuit. Other combinations are (a) the crystal unit frequency-temperature characteristics such as temperature coefficient and susceptibility to thermal shock, the oven temperature, and the degree of oven temperature control; (b) the crystal unit frequency-current characteristic, the oscillator current level and the oscillator current control. Care must be taken to see that no combination contributes more than about five parts in 10^{11} frequency change if the over-all design is to be stable to one part in 10^{10}.

In the design of AT-cut high-frequency shear mode crystal units the following two facts must be considered: (a) the Q of the quartz itself, at normal temperatures, increases as the frequency of operation is decreased[4] and (b) the lowest frequency at which the crystal unit can be operated without significant external losses is severely restricted by the availability of sufficiently large quartz plates. Since circuit phase stability is likely to be better at lower frequencies where the Q of quartz is higher, there is a definite advantage in the use of large quartz plates.

To achieve a Q limited only by the quartz itself, other causes of energy dissipation must be reduced to a negligible point. The following have been found effective:

(a) evacuate the enclosure, thus removing air damping;

(b) choose the size and shape of the quartz plate so that the edge is quiescent, thus eliminating energy loss through the edge and the mounting structure;

(c) polish the crystal plate major surfaces, thus removing minor imperfections that can dissipate energy in the active portion of the quartz plate (an improvement in Q of about 10 per cent can be achieved).

Once condition (b) has been met, the Q cannot be increased at a given frequency by changing the inductance of the unit, by using other modes of vibration or overtones, or by using electrodes of a different size. Under these conditions, the ratio of L_1 to R_1, and thus the Q, in the equivalent electrical circuit, Fig. 1, has been found to remain essentially constant, subject only to the three operations enumerated above. This is, of course, reasonable, since there is no change in the source of energy dissipation.

The inductance can be selected, therefore, to operate at an optimum impedance for a better match of crystal unit to circuit. An optimum impedance may be achieved by using an overtone mode of vibration (reversal of phase in the thickness direction). For a given frequency, an overtone mode unit requires a thicker quartz plate (desirable for frequency stability) and has values of L_1 and R_1 of Fig. 1 that are larger by the cube of the overtone employed. The impedance can also be raised by using other modes of vibration that are permitted by reversals in phase along the length or width of the quartz plate, or by parallel field excitation.[5] These methods, however, are less desirable than the use of the harmonically related overtone mode, since they do not permit the desirable increase in the thickness or frequency determining dimension.

Employing these principles, the crystal unit design proceeded about as follows:

i. The largest practical quartz plate, in view of the quartz supply and probable demand, was selected (30 mm diameter).

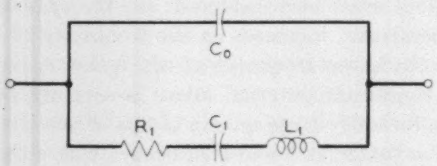


Fig. 1 — Equivalent circuit of crystal unit in vicinity of its operating frequency.

ii. The lowest frequency that could be used without energy loss at the edge of the quartz plate was determined (2.5 mc).

iii. The highest overtone (which is also the greatest thickness and highest $L_1 C_1$ ratio) that would allow a practical adjustment tolerance was chosen (fifth overtone and one part per 10^7 frequency adjustment tolerance).

The resulting crystal plate is believed to have the highest Q, the lowest frequency and the best impedance level that can be obtained from a 30-mm quartz plate in which the edge and mounting structure are not part of the mechanically vibrating (frequency-determining) part of the crystal unit.

III. DEVELOPMENT OF DESIGN AND PROCESSING TECHNIQUES

3.1 *Experimental Determination of Quartz Plate Size and Contour*

In a study[4] to determine the optimum contour and overtone for AT-cut, plated, 12.5-mm-diameter quartz plates, a series of measurements were made on several quartz blanks of different thicknesses, resonant at approximately 0.7, 1, 3 and 5 mc. Progressive contours from flat to the maximum permitted by the individual blank thickness were used. When these data were correlated, it became evident (a) that the maximum Q obtainable was an inverse function of frequency and (b) that there was a lower limit of frequency below which the Q fell off and became erratic regardless of contour. In other experiments a variation in electrode thickness from 700 to 2100 angstroms at 5 mc failed to show any effect on Q. Likewise, carefully polished quartz surfaces did not show more than a 10 per cent improvement in Q over that of carefully lapped and etched plates. Data taken at 10 mc on crystal units having quartz plates vibrating in the third, fifth and ninth overtone indicated the same maximum Q. This represents a 3-to-1 difference in quartz plate thickness and a 27-to-1 difference in the equivalent electrical inductance and resistance. Data were also taken on larger plates and on higher-frequency plates. The Q data from these tests are summarized on Fig. 2, which shows the most probable room temperature value for the internal friction of quartz, ranging from 15×10^6 at 1 mc to 0.15×10^6 at 100 mc, and the frequency limitation for 15-, 30- and 90-mm diameter plates. Therefore, with 30 mm having been chosen as the largest practical size for the quartz blank, the operating frequency of 2.5 mc is determined.

A chart relating optimum contour to overtone and frequency for AT-cut half-inch plano-convex plates can be found in an earlier paper by the author.[6] By linear enlargement or reduction of the dimensions, the ap-

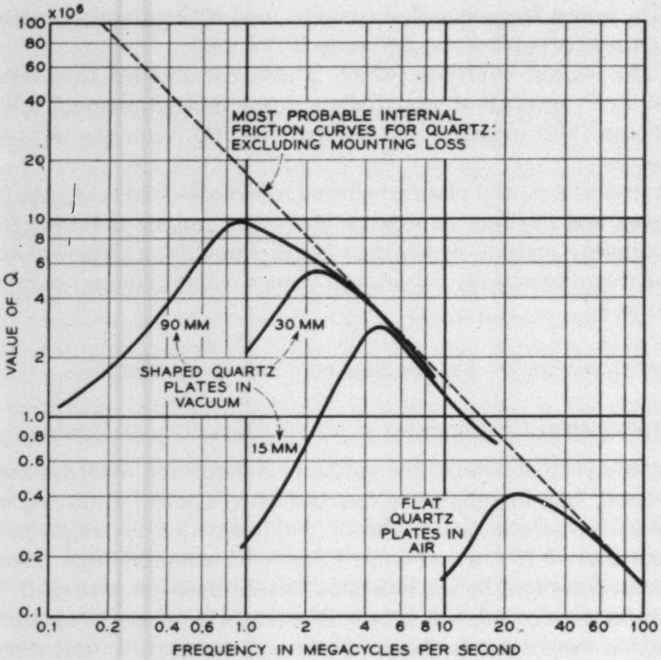


Fig. 2 — Value of Q vs. frequency for properly shaped quartz plates, 15, 30 and 90 mm in diameter. The value of Q is independent of the overtone mode of operation.

proximate contour for larger or smaller blanks can be determined, indicating a plano-convex contour 4 inches in radius for the 30-mm diameter, fifth overtone, 2.5-mc quartz plate.

It is well known[2] that a nodal plane exists that is centrally located between the faces of a quartz plate vibrating in thickness shear. For this reason, many AT-cut crystal units are designed with a double convex contour, with the mounting points on or near the nodal plane. It has been found, however, that, when the frequency, size and contour are chosen to produce the maximum Q, a plano-convex shape may be used with no loss in Q, and with great benefit in temperature-coefficient control and general handling during fabircation.

The final dimension to be determined, the thickness, was chosen to provide the correct impedance level for minimizing the effects of lead wire capacitance and circuit variations. A thickness of 3.4 mm was chosen, permitting operation on the fifth harmonically related overtone with a series-resonant resistance of 55 ohms and an inductance of 19.5 henries.

3.2 *Experimental Development of Quartz Polishing Methods*

The benefits derived from the use of polished quartz plates are improved electrical performance, particularly frequency stability at low current levels, and reduced frequency aging. The surface is not only more easily cleaned, since there are no scratches and fissures to trap contaminants, but the surface area is greatly reduced, requiring less gold for a conducting electrode and reducing the effects of residual contaminants.

The polishing techniques that have been developed differ in many respects from those used in the surface finishing of glass lenses. It is customary in polishing glass lenses to use a carefully prepared pitch lap and rouge, or its equivalent. Since pitch is a brittle material, close control over the curvature can be maintained. Furthermore, small scratches resulting from unavoidable foreign particles are reduced by the use of sufficient pressure to cause local melting and flow of the glass.

Such methods have not been found suitable for contoured quartz plates, nor are they necessary. The curvature is not critical, so there is no need for a brittle lap. Quartz is harder and has a higher melting point than glass and is crystalline in form, and any melting or scratch removal is both undesirable and difficult. A soft material such as an asphalt or cork mixture has proved better for the lap, since it can yield under pressure to give a uniform polish and can absorb foreign particles that would otherwise scratch the surface. Fig. 3 shows a polishing machine using two Trojan automatic bowl-feed sphere polishers. The polishing bowl has been covered with a $\frac{1}{16}$ inch sheet of cork and rubber (Corprene, Armstrong Cork Co.). Barnsite, a form of cerium oxide, is used as the polishing agent.

If polishing time is to be kept within practical limits, care must be given to surface preparation prior to polishing. There are two requirements: first, that surface penetration be small and second, that good thickness control be maintained, since final polish must occur at a thickness determined by the resonant frequency of the quartz blank. Both of these requirements have been met by the use of a resinoid-bonded diamond wheel to generate the convex surface. The apparatus is similar to diamond curve generators used in the lens inudstry, with the following exceptions: (a) a vacuum chuck is used to precisely hold the quartz blank; (b) a 3-inch-diameter, 180-mesh, resinoid-bonded diamond wheel is used and (c) a positive mechanical feed at 0.012-inch per minute is used. Thickness can be controlled within 0.01 mm, and the time of grinding is less than 3 minutes. The penetration is less than 20 microns, and may be removed by lapping for a few minutes with a cast-iron lap and emery mixture, followed by 5 to 10 minutes of polishing.

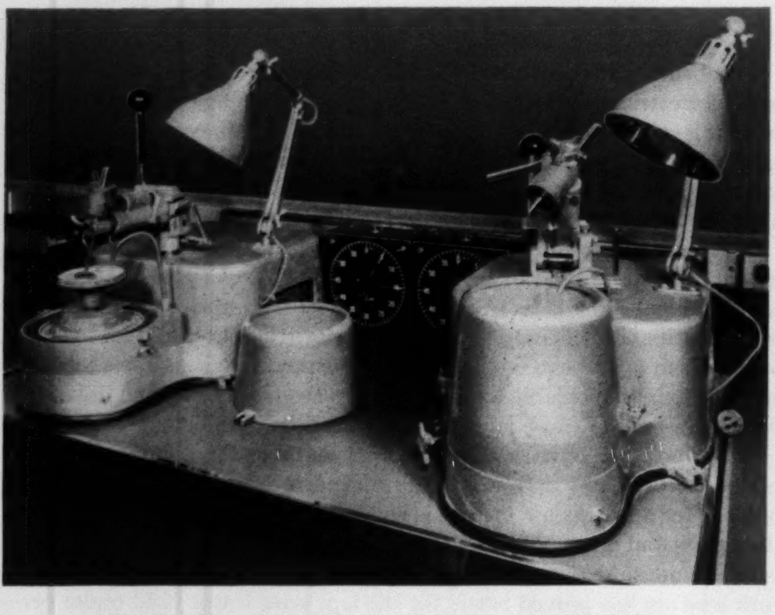


Fig. 3 — Equipment used to polish quartz crystal plates.

Three experimental procedures were developed in connection with the study of polishing techniques and the resultant quartz crystal surfaces:

First, a 200-power microsocpe was equipped with a dark field condenser, which clearly delineated scratches and cracks in the quartz surface. Fig. 4 is an enlargement of a picture taken through this microscope of what appeared to the unaided eye to be a well-polished blank. Since the blank is curved, all portions are not in focus in the picture. By refining the polishing technique, i.e., choosing best pressure, stroke and time, as well as the best preparation, surfaces that appeared clear by this inspection were consistently produced with 5 to 10 minutes of polishing.

Second, the spread of values for the Bragg angle of the $01\bar{1}$ face was measured, using a double-crystal goniometer.[7] This apparatus, shown in Fig. 5, is used principally for orientation measurements connected with the temperature coefficient. However, by using the same refined polishing techniques for the reference crystal, extremely sharp curves were obtained when the amplitude of the reflected X-ray beam was plotted against orientation. Fig. 6 shows typical results for quartz plates at various stages of polish. In particular, the use of etching to remove strained, slightly misoriented material is shown to be unnecessary sub-

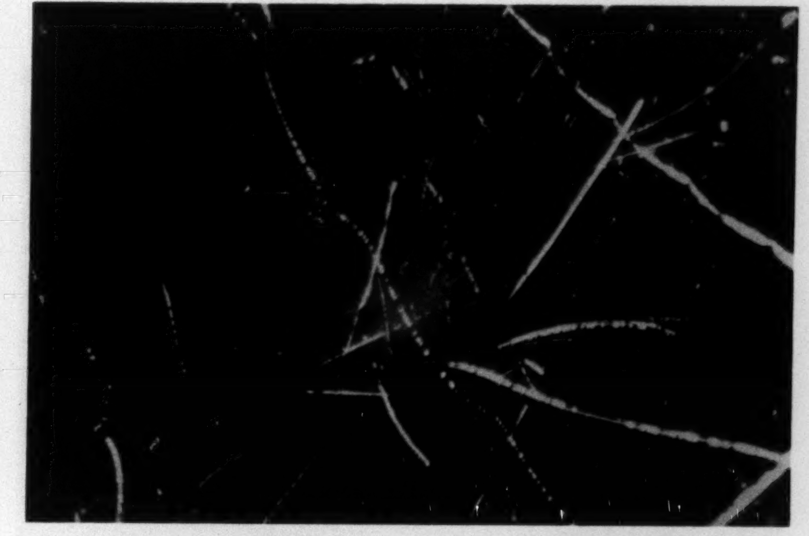


Fig. 4 — Dark-field photomicrograph of polished quartz plate; magnified 1000×.

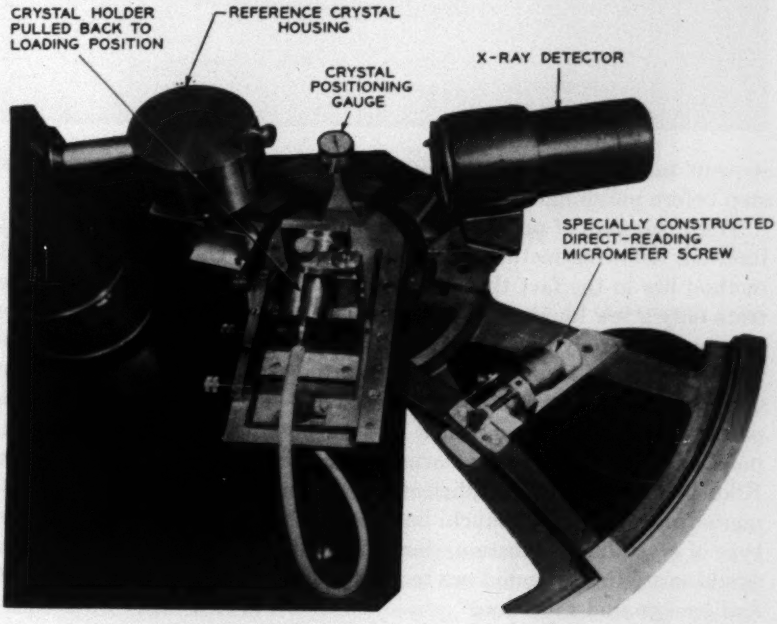

CRYSTAL HOLDER PULLED BACK TO LOADING POSITION
REFERENCE CRYSTAL HOUSING
CRYSTAL POSITIONING GAUGE
X-RAY DETECTOR
SPECIALLY CONSTRUCTED DIRECT-READING MICROMETER SCREW

Fig. 5 — Double-crystal goniometer used for orientation measurements.

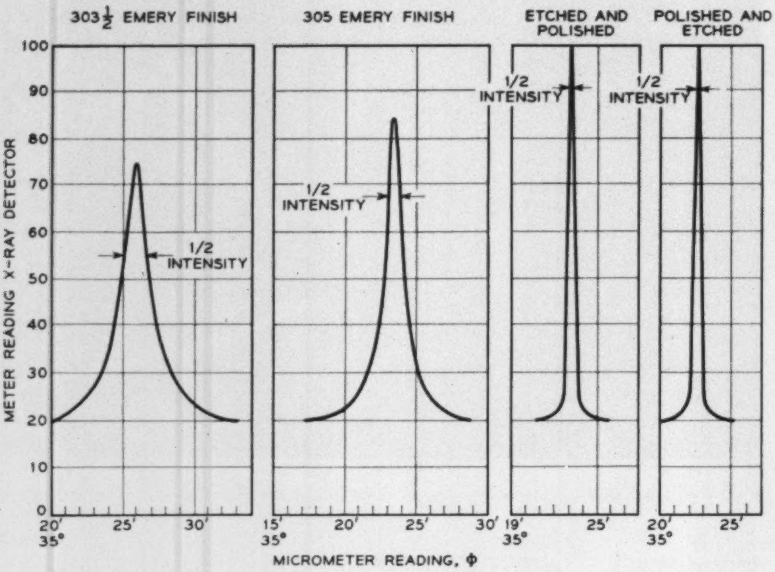


Fig. 6 — Typical response of the double-crystal goniometer for quartz plates at various stages of polish.

sequent to polishing, and therefore this is better performed as the last step before polishing.

Third, samples of polished quartz plates were studied by electron diffraction, following methods outlined by Arnold.[8] The advantage of this method lies in the fact that a beam of fast electrons (50 kv) will penetrate only a few hundred angstrom units before diffraction takes place, thus involving only the first few surface layers of the quartz plate. Should the formation of a misoriented or amorphous surface layer result from the polishing processes, it would be evident in the resulting diffraction pattern. Fig. 7 shows one such pattern obtained from a quartz plate polished using asphalt and barnsite. The lines observed are known as Kikuchi lines, and it is sufficient for the purpose of this discussion to quote from Arnold:[8] "Kikuchi line patterns are indicative of the highest type of crystalline perfection, since the slightest distortion of the crystal would cause the Kikuchi lines to spread out and become lost in the general background radiation."

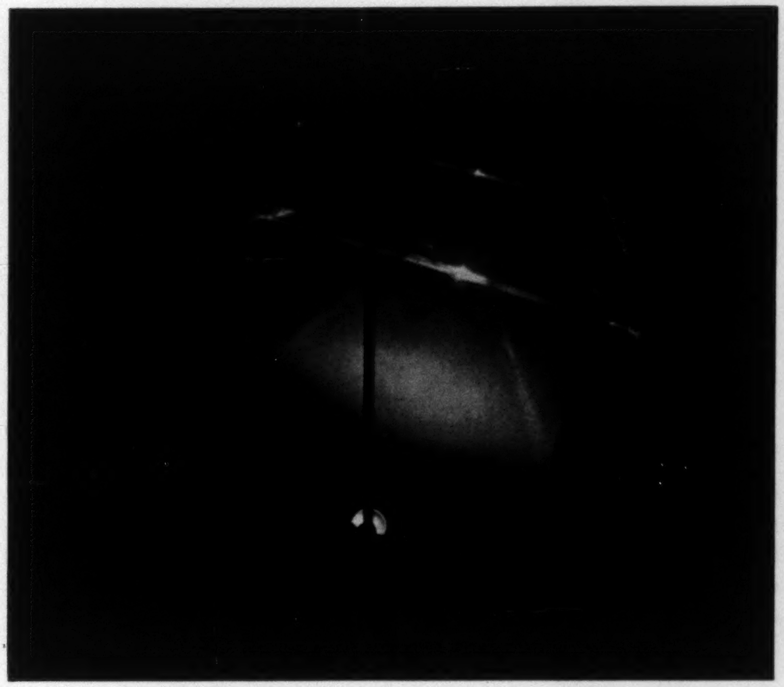


Fig. 7 — Electron diffraction pattern, showing Kikuchi lines.

3.3 *Studies of Correlation Between X-Ray Orientation and Temperature Coefficient*

The relationship between the resonant frequency of the quartz resonator, f, and temperature, t, can be expressed by

$$\frac{f}{f_i} = 1 - (24 \times 10^{-11})(t_0 - t_i)^2(t - t_i) + (8 \times 10^{-11})(t - t_i)^3, \quad (1)$$

where

f_i = frequency at inflection temperature, 27°C,
t_i = temperature of inflection point,
t_0 = temperature at which the temperature coefficient is zero.

The value of $t_0 - t_i$, which establishes the temperature at which the temperature coefficient of frequency is zero, is a function of the orienta-

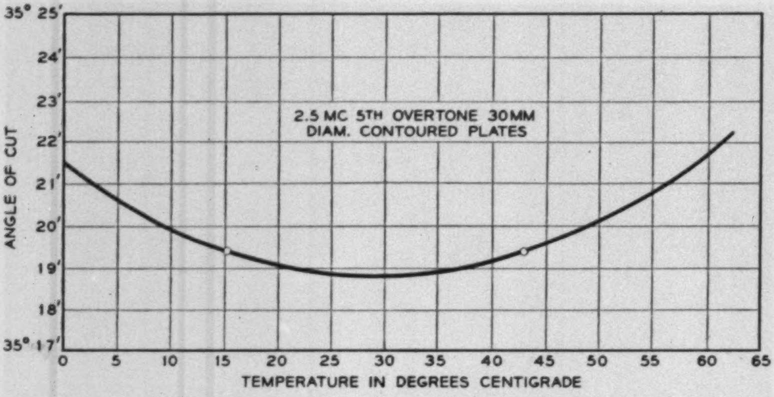


Fig. 8 — Temperature at which the temperature coefficient of frequency is zero vs. the crystal plate orientation about the x axis.

tion of the quartz plate with respect to its crystallographic axes, in particular the rotation about the x axis, Φ. Fig. 8 shows the best determination to date of this relationship. By plotting f versus t for various values of t_0 in (1), a family of curves is produced, as shown in Fig. 9. A close control over the angle of cut not only permits specification of an operating temperature, t_0, near room temperature, but also provides a much better temperature coefficient in the vicinity of t_0. This makes it less necessary to be concerned about an exact determination of t_0 or about a small shift in the oven control temperature. Determination of the angle to a few tenths of a minute of arc is very desirable, along with a close correlation between the measured angle and the observed temperature coefficient. The problems are related to the following requirements:

(a) an X-ray beam capable of resolving 0.1′ of arc;

(b) a crystalline surface sufficiently free from misoriented quartz;

(c) a method of defining the plane that controls the temperature coefficient;

(d) sufficiently accurate jigs and fixtures.

The double-crystal X-ray goniometer was shown in Fig. 5; it is a modified General Electric XRD1. Requirement (a) above is fulfilled by the use of a polished reference crystal from which a well-defined beam is reflected.[7] A quartz surface prepared as described above (Section 3.2) is more than adequate to meet requirement (b).

The nodal plane of Section 3.1, which controls the temperature coefficient, is, of course, physically inaccessible. However, an adequate surface [requirement (c)], from which to determine the temperature coefficient

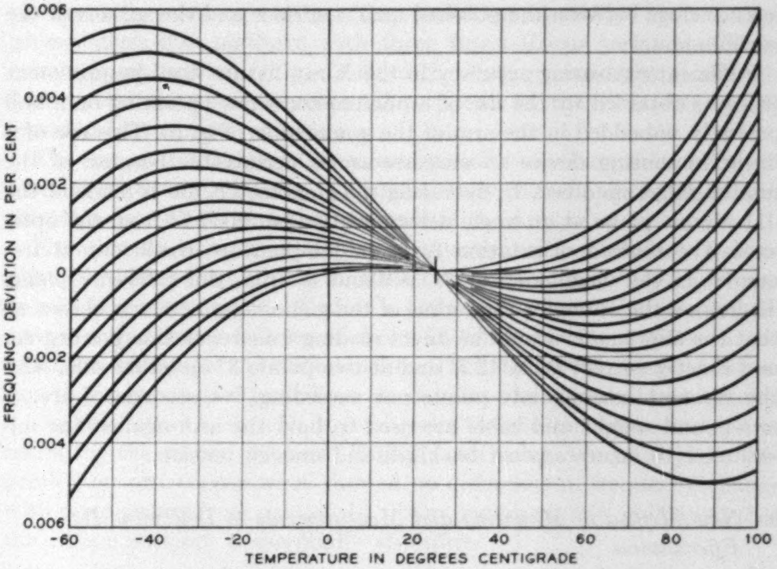


Fig. 9 — Frequency vs. temperature for different values of t_0, the temperature of zero temperature coefficient.

in contoured quartz plates, may be obtained by plano-convex shaping. The orientation of the convex side has almost no effect on the temperature coefficient, since a slight tilt of this surface with respect to the flat side only shifts the point of greatest thickness a little off center. The 16-mm diameter electrodes more than cover the actively vibrating portion of the crystal unit, and there is no measurbale effect on performance.

The principal problem in measuring the effective orientation of the flat side — that of physically defining the surface — resolves itself into a choice between two methods of securing the crystal to the goniometer table: (a) the use of three reference points and (b) the use of a reference plane. If irregularities exist in the quartz surface, there is a possibility that one or more points will not be representative of the controlling surface at the center of the plate. Likewise, when a reference plane is used, the presence of dust or contamination or a slight departure from flatness can shift the orientation. The work described in this article was done using a vacuum chuck with a polished reference surface. Assurance of cleanliness and reasonable flatness was obtained by observing interference rings between the polished quartz surface and the surface of the vacuum chuck.

Sufficient measuring accuracy in the X-ray fixture itself [requirement (d)] was obtained by the use of a micrometer screw operating on a ball precisely imbedded in the arm of the goniometer (Fig. 5). This use of a linear measuring device to measure arc is permissible because of the limited range involved. In operating the goniometer, use is made of the $01\bar{1}$ atomic plane at an angle determined to be 38°12.7′ from the optic axis. The desired orientation for zero temperature coefficient of frequency in the vicinity of 35°20′ is about 3° from this reference plane. Therefore, the value for the radius of the goniometer arm was chosen so that the micrometer would be direct reading (one revolution per degree) and exactly correct at 38°12.7′ and at two points 3° on either side, with the error at intermediate points not exceeding five seconds of arc. A two-pound weight and cable are used to hold the arm against the micrometer to insure against backlash and uneven tension.

3.4 *New Method of Mounting and Measurements to Determine Its Effectiveness*

A new mounting structure, Fig. 10, was devised for the 2.5-mc crystal unit in order to provide a support that was rigid yet free from the effects

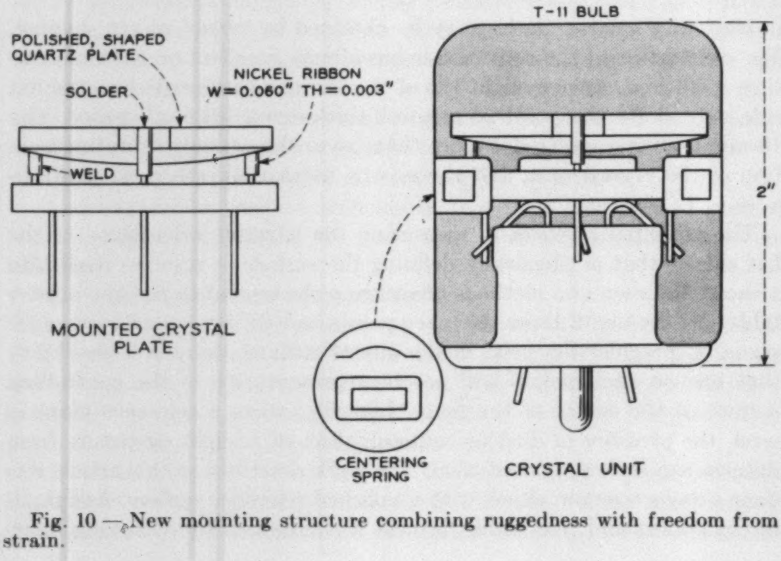


Fig. 10 — New mounting structure combining ruggedness with freedom from strain.

of thermally induced strains. The mounting assembly consists of a pressed-glass disc platform with three fused Kovar terminals. These terminals are welded on one side to the stem press of the glass bulb. The crystal plate is fastened to the terminals on the other side by ribbon-shaped elements of nickel.

The use of the three-ribbon mount permits relatively free radial expansion of the crystal plate, while adequately restraining the plate from translation or rotation during mechanical shock.

Experiments using crystal plates mounted with 0.050-inch rods in place of ribbons have shown that the time for the frequency to recover to within a few parts in 10^8 after a large temperature change (such as an oven shutdown) is reduced from 12 hours for the 0.050-inch rod mount to 2 hours for the ribbon mount. The time for frequency stabilization to about one part per billion per day likewise appears to be affected by residual strains, since it is two weeks for the rod support and two hours for the ribbon support. Experiments using a crystal plate suspended on soft copper wires showed no difference in frequency change with temperature from that of the ribbon-mounted unit, indicating that the ribbon support is essentially strainfree.

3.5 *Procedures Used in Forming Electrodes*

Electrodes are required in order to couple piezoelectrically to the quartz plate. From the standpoint of stability of the mechanical resonance, such electrodes would be best placed outside of the crystal plate enclosure. However, electrical considerations, such as the value and stability of the static capacitance, require that the electrode be an integral part of the vibrating quartz plate.

Gold is used as the electrode material because of proven characteristics such as ease of deposition, good electrical conductivity, softness, resistance to corrosion and good stability with time. Every effort is made to insure that the gold film, which is formed by evaporation under vacuum, is pure, soft and dense. To be sure, the handling properties of plated crystal units during fabrication are enhanced when certain impurities are present. Zinc and aluminum are effective in making the gold electrode relatively hard, adherent and scratch-resistant. Such electrodes are not, however, best for applications where the highest precision is desired. Experiments have shown that small amounts of impurities (<1 per cent) contribute to frequency-aging, probably through migration of one metal through the other, and that the superior adherence contributes to frequency instability through strains set up at the gold quartz interface.

In order to eliminate surface contamination, the vacuum system employed was specially designed, using oil-free bakable solenoid-operated valves and liquid-nitrogen traps. All vacuum baking to outgas the surface and to provide a hot substrate is done with large-area, relatively low-temperature conducting-glass plates rather than with open filaments. Up to five quartz plates are mounted vertically in the plating chamber, and electrodes are formed simulanteously on both sides by evaporating gold from eight small tungsten heaters, which are placed to assure even distribution. The apparatus is shown on Fig. 11. A gold electrode 16 mm in diameter and about 700 angstroms thick has proved adequate for this application.

3.6 *Frequency-Adjustment Technique*

The exact frequency desired from a crystal-controlled oscillator is obtained partly by controlling the natural resonant frequency of the crystal resonator during fabrication and partly by selecting or adjusting circuit elements in the oscillating loop. The adjustment of the natural resonance during fabrication of the quartz plate is simplified when the

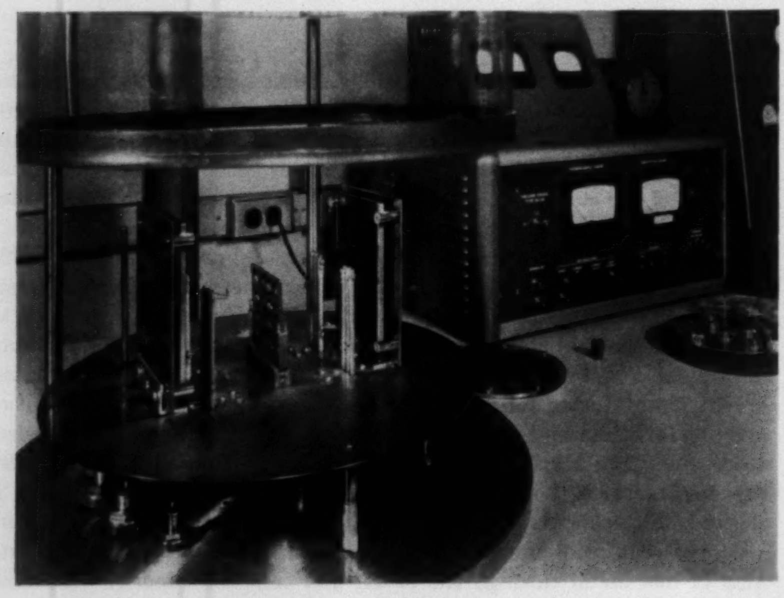


Fig. 11 — Apparatus used to form gold electrodes in vacuum on quartz plates.

control range is large. However, an upper limit to the control range is set by the probable stability of the controlling circuit element, usually a series capacitor. The slope of the crystal unit reactance with frequency is about 0.7 ohms for one part in 10^9 frequency change, as seen in Fig. 13. For example, an assumed instability of a series capacitor of only 0.01 per cent would require a value of 100$\mu\mu$f or larger to limit the frequency change to one part per 10^{10}. Under these conditions, practical limitations on size of the capacitor would limit the adjustable range to a few parts in 10^7.

Adjustment of the resonant frequency of the quartz plate to this degree is accomplished by adding gold to the electrode surface while the crystal plate is in oscillation, making use of the vacuum evaporation apparatus described above. The sequence of operations is as follows: (a) The exact frequency change desired is measured under final use conditions — that is, at operating temperature and proper circuit adjustment. (b) The crystal unit is placed in the vacuum chamber and gold deposition initiated. (c) The frequency *change* is monitored and controlled by continuous frequency measurement during deposition.

This method will usually result in finished crystal units not more than five parts per 10^7 from nominal frequency. The small error in frequency results principally from subsequent glass sealing operations and the cleaning effect of a final vacuum bake. When sufficient numbers of crystal units are processed in series, closer tolerances can be obtained by a method of compensation that uses measurements of finished units to provide information for the frequency adjustment of subsequent units

3.7 *Hermetic Seal Techniques*

Measurements of frequency aging of crystal units in both metal and glass enclosures[9] have shown the superiority of glass enclosures, probably because glass can be more effectively outgassed and cleaned at the temperatures involved.

The crystal plate should not be exposed to high temperatures, both to prevent a shift in resonant frequency and to avoid damage to the mounting attachment at the quartz plate. For this reason, a flared stem assembly and a close-tolerance baffle plate (also used as a support) are employed to keep the high temperatures involved in glass sealing away from the quartz plate. For the same reason, the seal is accomplished quickly with a minimum of glass annealing.

Following the stem-to-bulb seal, the unit is evacuated and baked for six hours at 140°C. The optimum length of time has been established experimentally, and is a function of the vacuum system design. With the new vacuum system described above, using oil-free valves and specially designed liquid nitrogen traps, it has been found that a six-hour bake can be used to good effect.

Following the baking, and with the vacuum at about 10^{-6} mm of mercury, the glass tabulation is sealed by means of a small flame.

IV. PROPERTIES OF THE QUARTZ RESONATOR

4.1 *The Crystal Unit as a Circuit Element*

Table I lists the electrical properties of the new 2.5-mc crystal unit, and the equivalent circuit of the crystal unit in the vicinity of its operating frequency was shown in Fig. 1. The capacity in the upper branch represents the static capacitance of the crystal and its holder. The lower branch represents the electrical equivalence of the mechanical resonance of the crystal, which has an impedance approximately given by

$$Z_1 = R_1 + j2\omega L_1 \frac{\Delta f}{f_r}, \tag{2}$$

where Δf is the difference between the operating frequency and the crystal series resonant frequency, f_r. The total impedance of the crystal, then, is

$$Z_c = \frac{Z_1 Z_0}{Z_1 + Z_0}. \tag{3}$$

This simplifies to

$$Z_c = R_e + jX_e = \frac{R_1}{1 - 2\dfrac{C_0}{C_1}\dfrac{\Delta f}{f_r}} = \frac{j2\omega L_1 \dfrac{\Delta f}{f_r}}{1 - 2\dfrac{C_0}{C_1}\dfrac{\Delta f}{f_r}} \tag{4}$$

when one uses the fact that the magnitude of the impedance Z_1 is much smaller (at the operating frequency) than the magnitude of the impedance Z_0.

TABLE I

Series resonant resistance, R_1	65 ± 10 ohms
Inductance, L_1	19.5 henries
Dynamic capacitance, C_1	0.00021 μμf
Q	4×10^6
Static capacitance, C_0	4.0 μμf
$r = c_0/c_1$	19,000
Nominal capacitance for operation at standard frequency	50 μμf
Manufacturing tolerance on frequency	±6 pp 10^7

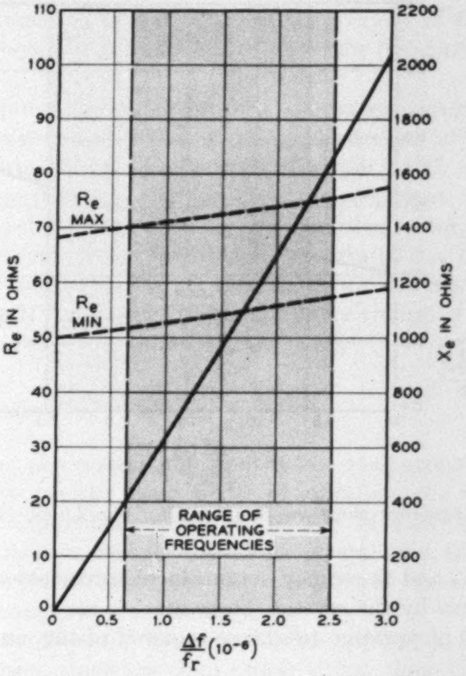


Fig. 12 — Effective resistance and reactance vs. frequency for the 2.5-mc crystal unit.

The first term of (4) is the effective resistance, R_e, and the second term the effective reactance, X_e, of the crystal. These are plotted versus the fractional frequency deviation from crystal resonance, $\Delta f/f_r$, in Fig. 12. The range of operating frequencies shown in the figure is based on the crystal manufacturing tolerances and the expected total aging.

The sensitivity of the oscillating frequency to changes in the reactance of the circuitry associated with the crystal depends on the "stiffness" or reactance slope of the crystal at the operating frequency. This is obtained by differentiating X_e with respect to fractional frequency deviations:

$$\frac{dX_e}{d(\Delta f/f_r)} = \frac{2\omega L_1}{\left(1 - 2\frac{C_0}{C_1}\frac{\Delta f}{f_r}\right)^2}. \tag{5}$$

Equation (5) is plotted as a function of f/f_r in Fig. 13.

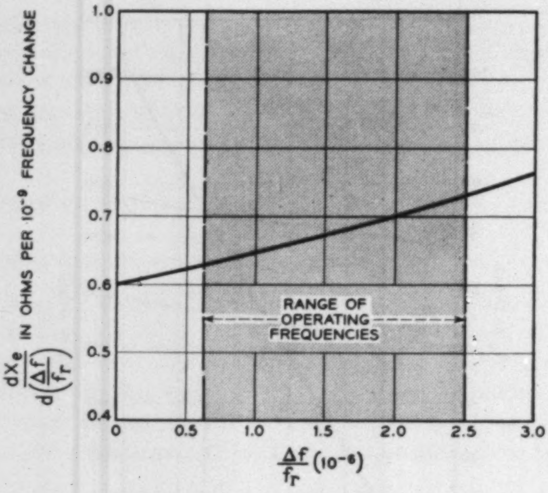


Fig. 13 — Reactance slope vs. frequency for the 2.5-mc crystal unit.

From Figs. 12 and 13 we may obtain the requirements imposed on the oscillating circuit by the crystal. These are:

i. The range of negative resistance required of the circuit is −51 to −75 ohms.

ii. The negative reactance of the circuit should be adjustable from 400 to 1700 ohms.

iii. The total negative reactance of the circuit should be stable to better than 0.1 ohm for a frequency stability of one part in 10^{10}.

Other requirements imposed by the crystal on the circuit are:

iv. The crystal current should be stabilized at about 70 microamperes to a constancy of 1 db.

v. The circuit should contain elements to prevent oscillation at unwanted crystal modes of resonance, in particular the third overtone frequency near 1.5 mc.

4.2 *Temperature Coefficient of Frequency*

The relationship between orientation of the quartz plate with respect to its crystallographic axes and the temperature of the zero temperature coefficient is shown on Fig. 8. By maintaining the angle to 35°20′ ± 1′ the temperature coefficient will go through zero at a temperature, t_0, which lies between 42°C and 57°C. In the vicinity of t_0, the relative deviation of frequency at temperature t from the frequency at t_0 is given by (1).

In order to prevent possible temperature-control aging of 0.1°C from causing a frequency change of more than one part per 10^{10}, the temperature of the thermostat must agree with t_0 within 0.1°C.

It can be seen that the actual temperature coefficient, which is better than one part per 10^9 per degree, is not a limiting factor in any reasonable oven construction. On the other hand, strains due to temperature changes in the quartz itself are a limiting factor and a temporary shift of one part per 10^{10} will occur if a temperature change of 5 millidegrees per hour is maintained for 10 minutes or more.

4.3 *Current Coefficient of Frequency*

The frequency of a crystal unit depends to a small extent on the crystal current. If uncoupled to other modes of vibration, the relationship at low currents is approximately $\Delta f/f = Di^2$. Fig. 14 shows a typical curve of frequency versus current for the 2.5-mc crystal unit. In order to keep the current coefficient below one part per 10^9 per db, currents of less than 100 microamperes are necessary.

The current coefficient of frequency in this application is not believed due to dissipation, since the total power is less than 10^{-7} calorie per second, and also because the effect is nearly instantaneous. The most likely explanation is that the elastic constant varies with strain; that is, Hooke's Law is really not obeyed. Further studies indicate that the frequency change is a function of the amplitude of the strains due to oscillation, and that it is independent of Q and frequency.

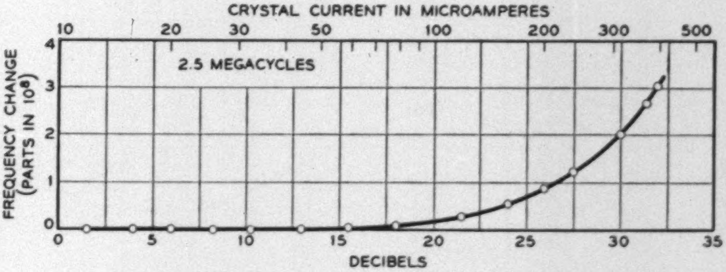


Fig. 14 — Frequency vs. crystal current for the 2.5-mc crystal unit.

4.4 *Mechanical Stability*

The crystal plate will withstand a static load of 2 lbs (200 g's) in any direction without any apparent movement with respect to its mounting platform. The mounting plate is in turn anchored to the glass bulb by three peripheral springs and the three nickel wires in the glass stem press. Severe shock, such as a four-inch drop, will dislodge the platform, and should be avoided. Normal shipping, however, should have no effect on the crystal unit properties. Similarly mounted 5-mc crystal units have withstood 10g vibration to 2000 cycles with no permanent frequency change greater than one part per 10^9.

There is an orientation effect on frequency caused by gravity-induced strains, as shown on Fig. 15. The preferred orientation is with the unit installed with the odd mounting ribbon vertical, which will allow a ±20° tilt without affecting the frequency more than one part per 10^{10}.

4.5 *Frequency Stabilization and Aging*

It is customary to differentiate between the rapid frequency drift associated with initial operation of a frequency standard, here called stabilization, and the slower frequency drift known as aging. Whereas the former will have become negligible after a few weeks or months, the latter can extend over several years.

Naturally, the drift should be as small as possible. If it cannot be avoided, it should be a simple function of time, to permit extrapolation.

No uniform result has been obtained in the initial stabilization of the quartz resonators. Evidence suggests[9] that the frequency change is due principally to a transfer of mass to and from the quartz plate, initiated by a shift in temperature. The rate of transfer and the degree of permanence of the transfer will be a function of the vapor pressure of the

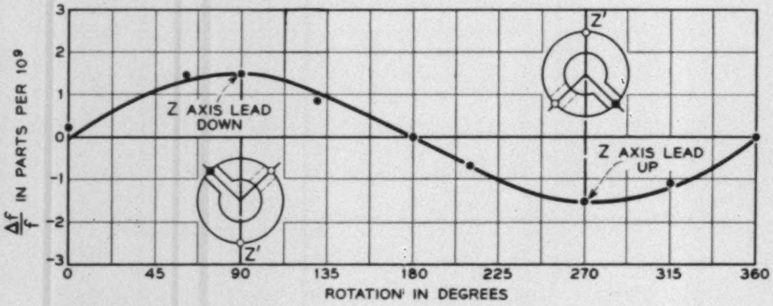


Fig. 15 — Effect of gravity-induced strains on frequency at 2.5 mc.

particular contaminant and the degree of adherence (which may be molecular, chemical or mechanical) between the contaminant and its substrate. Since the equilibrium reached at any given operating condition is not likely to repeat itself, the initial frequency drift cannot be accurately predicted. The magnitude of this drift is not more than a few parts in 10^8, and may be reduced as improved cleaning techniques in manufacture are developed. Use of a carefully controlled temperature cycle each time the oven is re-started can also reduce this initial uncertainty by as much as a factor of ten. In any case, the drift rate can be expected to decrease to about one part in 10^9 per month by the third month of operation.

The aging of new quartz resonators has not proved uniform, either. Operation at 50°C has, however, consistently shown less aging than operation at 75°C. Of five oscillators at 50° for which records have been kept, the rate varies from one to ten parts in 10^{10} per month. One such oscillator is used in connection with the National Bureau of Standards broadcast from Station WWV, and its frequency versus that of an atomicron at the station has been published.[10] Its aging rate after about 12 months of operation appears to be about two parts per 10^{10} per month. One oscillator operated at Bell Telephone Laboratories, Whippany, New Jersey, which has been monitored by use of a 60-kc broadcast by the National Bureau of Standards from Station KK2XEI in Boulder, Colorado and from MSF in Rugby, England, is shown on Fig. 16. This oscillator was considerably better than one part per 10^9 per month, even in the second month, but the indicated long time rate will be in the order

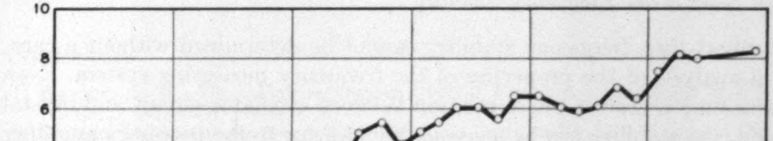


Fig. 16 — Frequency aging data, 2.5-mc crystal unit — KK2XEI received signal vs. Whippany frequency standard.

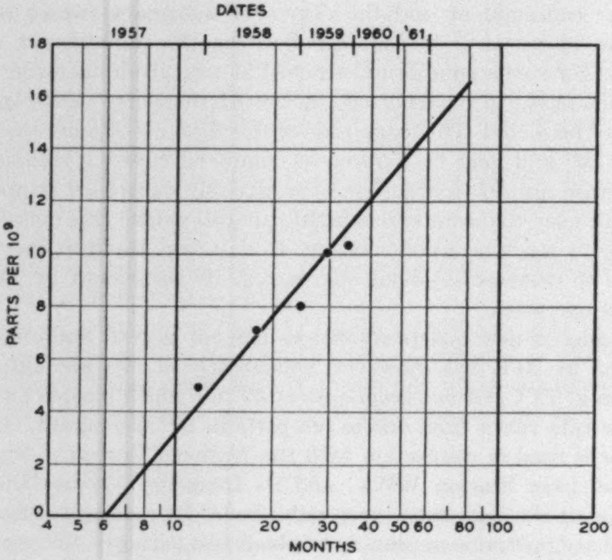


Fig. 17 — Frequency aging data shown to be a simple exponential curve.

of one part per 10^9 per year for some time to come. Fig. 17 shows that the frequency change is a simple exponential curve, and can be easily extrapolated.

4.6 *Short-Time Frequency Stability*

Short-time frequency stability cannot be determined without a careful analysis of the properties of the frequency measuring system. Even assuming a correct phase relation between oscillator circuit and crystal unit, the stability will be unavoidably lost due to the necessary amplifier and lines, and, ultimately, the measuring equipment itself. These phase distortions cannot be readily distinguished from true frequency variations.

Measurements have been made at various multiplier frequencies up to 1000 mc in an attempt to find the best conditions for measurement, and the following figures represent the results to date:

0.1-second averaging: two parts in 10^9;
1-second averaging: two parts in 10^{10};
10-second averaging: two parts in 10^{11}.

Since both oscillators contribute to the instability, we may assume that one oscillator is at least twice as good. That is, the mean relative deviation is one part in 10^{11} or better when the frequency is averaged for 10 seconds or longer.

V. CONCLUSION

The new 2.5-mc crystal units will make possible general-use oscillators characterized by high frequency stability, comparatively little aging, good linearity and uncomplicated design. Such standards compare favorably with atomic standards for periods up to one month or more, and have an advantage over atomic standards in that they may be set to an exact frequency and are more portable and rugged.

The crystal units are uniform in Q and frequency, and need not be specially selected. Although the use of a relatively high frequency and electrodes integral with the quartz plate might be questioned, experiments have demonstrated that the associated difficulties can be as easily dealt with as can those associated with a resonant mounting or isolated electrodes characteristic of low-frequency units. Various advantages accrue from the fact that only the center portion of the quartz plate and its pure gold electrodes determine the resonant frequency. Among these are exceptional stability under conditions of shock and vibration, and uniform and highly predictable electrical characteristics.

The development work leading to the design and fabrication of 2.5-mc crystal units and associated oscillators and ovens has been supported in part by development contracts with the Rome Air Development Center and the U. S. Army Signal Research and Development Laboratory.

REFERENCES

1. Warner, A. W., Ultra-Precise Quartz Crystal Frequency Standards, I.R.E. Trans., **I-7**, 1958, p. 185.
2. Heising, R. A., *Quartz Crystals for Electrical Circuits*, D. Van Nostrand Co., New York, 1946.
3. Awender, H., The Current Status of Quartz Crystal Development, Nachr. tech. Zeit., **11**, 1958, p. 225.
4. Bommel, H. E., Mason, W. P. and Warner, A. W., Dislocations, Relaxations and Anelasticity of Crystal Quartz, Phys. Rev., **102**, 1956, p. 64.
5. Bechmann, R., Improved High Precision Quartz Oscillators Using Parallel Field Excitation, Proc. I.R.E., **48**, 1960, p. 367.
6. Warner, A. W., High-Frequency Crystal Units for Primary Frequency Standards, Proc. I.R.E., **40**, 1952, p. 1030.
7. Bond, W. L., A Double-Crystal Goniometer, Proc. I.R.E., **38**, 1950, p. 866.
8. Arnold, G. W., Surface Structure of Quartz Crystals, Naval Research Lab., Report 4065, 1952.
9. Warner, A. W., Frequency Aging of High-Frequency Plated Crystal Units, Proc. I.R.E., **43**, 1955, p. 790.
10. National Bureau of Standards, WWV Standard Frequency Transmission, Proc. I.R.E., **47**, 1959, p. 1157; and in all issues following through January 1960 (except for September 17 through October 12, 1959).

Some Further Theory of Group Codes

By DAVID SLEPIAN

(Manuscript received April 5, 1960)

The notion of equivalence for group codes is explored in some detail. A dual for a code, and the sum and product of two or more codes, are defined. Properties of these constructs are investigated. Indecomposable codes are defined and are shown to be optimal in two different senses. Various classes of codes are enumerated.

INTRODUCTION

This paper is a collection of results on the theory of group error-correcting codes for use on binary channels. It investigates further certain topics introduced in an earlier paper[1] by the author. The reader will be assumed to be familiar with the contents of this earlier paper as well as with the general nature of the coding problem in information theory.

The evident trend to digital transmission systems has given rise in recent years to an increased interest in coding as a possible practical means of error control. Lacking an "explicit solution" to the coding problem in any real sense, many investigators have chosen in an *ad hoc* manner promising special classes of parity-check codes and have examined their properties. A large and useful literature of special codes has resulted.

The approach taken here is different. No special codes are examined; rather, we attempt to shed some additional light on the structure of the class of all group codes. Our original aim was to parametrize in some manner the various equivalence classes of group codes. If such a parametrization could be effected, one could then hope to express the error probability of a code in terms of the parameters, and possibly to see how to choose the parameters to obtain codes of small error probability. We have fallen far short of this goal.

The main results to be found in this paper are as follows. A natural dual for a group code is defined. For any two group codes, a product code and a sum code are defined and certain properties of these operations are investigated These operations have the important property of maintaining equivalence in the sense that if $\mathcal{A}$ and $\mathcal{A}'$ are equivalent group codes and $\mathcal{B}$ and $\mathcal{B}'$ are equivalent group codes, then $\mathcal{A} + \mathcal{B}$ is equivalent to $\mathcal{A}' + \mathcal{B}'$ and $\mathcal{A}\mathcal{B}$ is equivalent to $\mathcal{A}'\mathcal{B}'$. This result in turn leads to an arithmetic of equivalence classes of codes. The notion of an (additively) indecomposable equivalence class is introduced, and it is shown that an arbitrary equivalence class can be written in a unique manner as a sum of indecomposable equivalence classes. It is then shown that one can limit the search for best codes (with two commonly used meanings for "best") to the indecomposable equivalence classes. Enumeration formulae for the types of equivalence classes are given, and these formulae are evaluated for small values of the pertinent parameters.

In the interest of simplicity of exposition, we have restricted our attention to binary codes, although many of the results obtained hold for codes consisting of sequences of elements drawn from any finite field. Also, in an effort to make the paper available to as wide a class of readers as possible, we have carefully eschewed the specialized vocabulary of modern algebra,* although many of our results could be stated more succinctly in these terms. In addition, as an aid to the casual reader we adopt once more the format of Ref. 1: Part I contains definitions, examples and results; Part II contains additional theory and proofs of the less obvious assertions of Part I. The terminology of Ref. 1 is maintained with one exception: the word "code" is here used as a synonym for "alphabet," as has become accepted practice in the literature.

There is some overlap of material with that found in the paper of Fontaine and Peterson[2] which appeared after much of this work was done. In the interest of making this paper self-contained, we repeat some material that might have been quoted from that paper.

Part I — DEFINITIONS, EXAMPLES AND RESULTS

1.1 *Recall of Previous Paper*[1] *and Some New Definitions*

An (n,k)-alphabet, or (n,k)-code, is an unordered collection of 2^k distinct n-place binary sequences that forms an Abelian group under the operation of mod 2 addition of the sequences term by term. The elements of the group, that is, the n-place binary sequences, are also called "letters." We assume always in this paper that $n \geqq k > 0$.

We denote specific group codes by large script letters, $\mathcal{A}$, $\mathcal{B}$, etc. We denote the letters of $\mathcal{A}$ by A_1, A_2, etc., and the digits of a letter by lower-case Latin letters. Thus, for example, a particular letter of the (n,k)-

---

* In modern terminology, we are studying properties of subspaces of a finite dimensional linear vector space over a finite field.

code $\mathfrak{A}$ is the binary sequence $A_1 = (a_1, a_2, \cdots, a_n)$. It is frequently convenient to regard the letters A_1, A_2, etc. as n-dimensional vectors.

A particular (n,k)-code can be specified by listing its 2^k letters. It can also be specified by listing k of its generators, i.e., any k linearly independent letters of the code. These k generators can be displayed as a binary matrix of rank k, with k rows and n columns. The rows of the matrix are the generators of the code. Such a matrix will be called a *generator matrix* and will be denoted typically by the symbol Ω. When referring to different generator matrices of a specific code $\mathfrak{A}$, we shall write $\Omega_1(\mathfrak{A})$, $\Omega_2(\mathfrak{A})$, etc.

Many generator matrices correspond to the same code. The first generator can be chosen in $2^k - 1$ ways, since the all-zero sequence or identity, I, of the group code cannot serve as a generator. The second generator can be chosen in $2^k - 2^1$ ways. The third can be chosen in $2^k - 2^2$ ways, since the first two generators determine a group of order 2^2. Proceeding in this way, we find

$$\begin{aligned} M_k &= (2^k - 2^0)(2^k - 2^1)(2^k - 2^2) \cdots (2^k - 2^{k-1}) \\ &= 2^{k(k-1)/2}(2^k - 1)(2^{k-1} - 1)(2^{k-2} - 1) \cdots (3)(1) \end{aligned} \tag{1}$$

different generator matrices for a given (n,k)-code. Indeed, if Ω_1 and Ω_2 are generator matrices for the same code, then $\Omega_1 = g\Omega_2$, where g is a nonsingular $k \times k$ binary matrix and all operations implied in the matrix product $g\Omega_2$ are carried out mod 2. The collection of $k \times k$ nonsingular binary matrices forms a group under matrix multiplication (arthimetic mod 2) which we shall denote by G_k. G_k is of order M_k. [G_k is the general linear group of dimension k over a field of two elements, frequently denoted by $GL(k, 2)$.] If Ω is any generator matrix for an (n,k)-code, then, as g runs through G_k, $g\Omega$ gives the M_k distinct generator matrices associated with the code.

In all that follows we shall frequently omit the phrase "all arithmetic mod 2." It will generally be clear from the context whether the field in question is the reals, the complex numbers, or the two element field.

It was shown in Ref. 1 that every group code is a parity-check code and that every parity-check code is a group code. Let Λ be a binary matrix of $n - k = l$ rows and n columns and of rank l. Let λ_{ij} be the entry in the ith row and jth column of Λ, $i = 1, 2, \cdots, l$ and $j = 1, 2, \cdots, n$. The equations

$$\Lambda \tilde{A} = 0 \tag{2}$$

or

$$\sum_{j=1}^{n} \lambda_{ij} a_j = 0, \qquad i = 1, 2, \cdots, l,$$

where A is the binary row vector $A = (a_1, a_2, \cdots, a_n)$ and the tilde denotes transpose, have k linearly independent solutions, say $A_1, A_2, \cdots, A_k$. These k vectors can be taken as the generators of an (n,k)-code. Since every linear combination of the vectors $A_1, \cdots, A_k$ also satisfies (2), every generator matrix Ω of this (n,k)-code satisfies

$$\Lambda\tilde{\Omega} = 0.$$

The matrix Λ is called a *parity-check matrix* for the (n,k)-code.

A given (n,k)-code has many parity-check matrices. Indeed, if Λ is one such, so is $g\Lambda$ for every g contained in G_{n-k}. There are therefore M_{n-k} distinct parity-check matrices associated with a given (n,k)-code. We shall denote the different parity-check matrices of a specific (n,k)-code $\mathfrak{A}$ by $\Lambda_1(\mathfrak{A})$, $\Lambda_2(\mathfrak{A})$, etc.

1.2 *Equivalence*

As in Ref. 1, we define two (n,k)-codes to be equivalent if one can be obtained from the other by a fixed permutation of the places of every letter. The concept has been illustrated in Section 1.7 of Ref. 1. Equivalent (n,k)-codes have the same transmission properties over the binary symmetric channel.

We denote the fact that codes $\mathfrak{A}$ and $\mathfrak{B}$ are equivalent by the symbolism $\mathfrak{A} \cong \mathfrak{B}$. It is immediately established that this is a true equivalence relation; i.e., that $\mathfrak{A} \cong \mathfrak{A}$; that $\mathfrak{A} \cong \mathfrak{B}$ implies $\mathfrak{B} \cong \mathfrak{A}$; and that if $\mathfrak{A} \cong \mathfrak{B}$ and $\mathfrak{B} \cong \mathfrak{C}$, then $\mathfrak{A} \cong \mathfrak{C}$. The totality of (n,k)-codes can therefore be broken down into disjoint equivalence classes. We denote by $\hat{\mathfrak{A}}$ the equivalence class containing $\mathfrak{A}$.

This equivalence of codes induces an equivalence relation among the totality of possible generator matrices. Two such matrices, say Ω_1 and Ω_2, will be called equivalent (written $\Omega_1 \cong \Omega_2$) if there exists a g in G_k and an $n \times n$ permutation matrix σ such that $g\Omega_1\sigma = \Omega_2$. That is, two $k \times n$ Ω-matrices are equivalent if one can be obtained from the other by permuting columns and/or forming nonsingular linear combinations of the rows mod 2. Clearly, two equivalent Ω-matrices, when considered as generator matrices, give rise to equivalent codes. Equivalent codes have equivalent generator matrices.

The task of analyzing group codes would be greatly simplified if a canonical form could be found for each equivalence class of Ω-matrices. That is, for a given n and k, we should like to be able to write down one generator matrix from each equivalence class. This would provide a simple means of describing each of the essentially different (n,k)-codes. The number of equivalence classes of (n,k)-codes is very much smaller than the number of distinct (n,k)-codes. They are enumerated in Section 1.9. Here we present further only two results pertaining to equivalence.

Every $k \times n$ Ω-matrix is equivalent to an Ω-matrix whose first k rows and columns are the $k \times k$ unit matrix. That is, Ω is equivalent to the partitioned matrix $\Omega \cong (I_k \vdots M)$, where I_k is the $k \times k$ unit matrix and M is a matrix of k rows and $l = n - k$ columns.

An Ω-matrix with the above structure will be said to be in *M-form.* Unfortunately, two $k \times n$ Ω-matrices in M-form having different M-matrices (even apart from permutations of rows and columns) can be equivalent.

A second result is

Theorem 1: A necessary and sufficient condition for two $k \times n$ Ω-matrices to be equivalent is that their columns can be placed into a one-to-one correspondence that preserves mod 2 addition of the columns.

Examples: Let

$$\Omega_1 = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}, \qquad \Omega_2 = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix}.$$

Then $\Omega_1 \cong \Omega_2$, for if we denote the columns of Ω_1 by $u_1, u_2, \cdots, u_5$ and those of Ω_2 by $v_1, v_2, \cdots, v_5$ and establish the correspondence $u_1 \leftrightarrow v_3$, $u_2 \leftrightarrow v_5$, $u_3 \leftrightarrow v_2$, $u_4 \leftrightarrow v_1$, $u_5 \leftrightarrow v_4$, one sees that u_1, u_2, u_3 are independent as are v_3, v_5, v_2 and that the equations $u_4 = u_1 + u_2$ and $u_5 = u_1 + u_2 + u_3$ have the analogs $v_1 = v_3 + v_5$ and $v_4 = v_3 + v_5 + v_2$. Both Ω_1 and Ω_2 are equivalent to

$$\Omega_3 = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

The matrices Ω_1 and Ω_3 are both in M-form and are equivalent, although they have different M-matrices.

The preceding considerations of equivalence for Ω-matrices have their obvious analogs for parity-check matrices.

1.3 *Duality*

There is a natural duality between (n,k)-codes and (n,l)-codes, where $l = n - k$. In Ref. 1 it was noted that the two sets of codes are equinumerous. We elaborate further on this notion here.

In Section 1.1 it was remarked that every generator matrix $\Omega(\mathcal{A})$ for a given (n,k)-code $\mathcal{A}$ and every parity check matrix $\Lambda(\mathcal{A})$ for this code satisfies

$$\Lambda(\mathcal{A})\tilde{\Omega}(\mathcal{A}) = 0. \tag{3}$$

The transpose of this relation is

$$\Omega(\mathcal{A})\tilde{\Lambda}(\mathcal{A}) = 0.$$

Thus, *every parity check matrix* $\Lambda(\mathcal{A})$ *of an* (n,k)*-code* $\mathcal{A}$ *can be regarded as a generator matrix for a particular* (n,l)*-code hereafter called the dual of* $\mathcal{A}$ *and denoted* $\mathcal{A}^\dagger$. *Every generator matrix* $\Omega(\mathcal{A})$ *is a parity check matrix for* $\mathcal{A}^\dagger$.

The above can be regarded as defining $\mathcal{A}^\dagger$ by the relation

$$\Omega(\mathcal{A}^\dagger) = \Lambda(\mathcal{A}).$$

One immediately finds that

$$(\mathcal{A}^\dagger)^\dagger = \mathcal{A} \tag{4}$$

and that

$$\mathcal{A} \cong \mathcal{B} \quad \text{implies} \quad \mathcal{A}^\dagger \cong \mathcal{B}^\dagger. \tag{5}$$

The equivalence classes of (n,k)-codes can therefore be put in a natural way into one-to-one correspondence with the equivalence classes of (n,l)-codes:

$$\hat{\mathcal{A}} \text{ corresponds to } \widehat{\mathcal{A}^\dagger}.$$

It is convenient to define

$$(\hat{\mathcal{A}})^\dagger = \widehat{\mathcal{A}^\dagger}.$$

There is a simple way of passing from a $k \times n$ generator matrix Ω in M-form for a code in $\hat{\mathcal{A}}$ to a generator matrix Ω' in M-form for a code in $\hat{\mathcal{A}}^\dagger$. If $\Omega = (I_k \vdots M)$ defines a code in $\hat{\mathcal{A}}$, then $\Omega' = (I_l \vdots \tilde{M})$ defines a code in $\hat{\mathcal{A}}^\dagger$. Here $\tilde{M}$ is the transpose of M.

1.4 *The Sum of Two Codes*

Let $\mathcal{A}$ be an (n,k)-code and $\mathcal{B}$ be an (n',k')-code. We define a new code $\mathcal{C}$ by the partitioned generator matrix

$$\Omega(\mathcal{C}) = \begin{pmatrix} \Omega(\mathcal{A}) & \vdots & 0 \\ \cdots\cdots & \vdots & \cdots\cdots \\ 0 & \vdots & \Omega(\mathcal{B}) \end{pmatrix}. \tag{6}$$

The code $\mathcal{C}$ *is an* $(n + n', k + k')$*-code called the sum of* $\mathcal{A}$ *and* $\mathcal{B}$ *and we write* $\mathcal{C} = \mathcal{A} + \mathcal{B}$. It is easy to show that this is a valid definition and does not depend on the particular generator matrices chosen for $\mathcal{A}$ and $\mathcal{B}$.

If $\Lambda(\mathcal{A})$ and $\Lambda(\mathcal{B})$ are parity-check matrices for $\mathcal{A}$ and $\mathcal{B}$ respectively, then

$$\Lambda(\mathcal{C}) = \left(\begin{array}{c:c} \Lambda(\mathcal{A}) & 0 \\ \hdashline 0 & \Lambda(\mathcal{B}) \end{array}\right) \tag{7}$$

is a parity-check matrix for $\mathcal{C} = \mathcal{A} + \mathcal{B}$.

Transmission of a letter from $\mathcal{C}$ amounts to transmitting a letter from $\mathcal{A}$ followed by a letter from $\mathcal{B}$. Because of the independence of the noise on the channel from one transmitted digit to the next,* it follows at once that *if* $Q_1(\mathcal{A})$, $Q_1(\mathcal{B})$ *and* $Q_1(\mathcal{C})$ *(see Section 1.6, Ref. 1) are the probability of no error for codes* $\mathcal{A}$, $\mathcal{B}$ *and* $\mathcal{C} = \mathcal{A} + \mathcal{B}$ *respectively, then* $Q_1(\mathcal{C}) = Q_1(\mathcal{A})Q_1(\mathcal{B})$.

If $\mathcal{C} = \mathcal{A} + \mathcal{B}$, a generator matrix for $\mathcal{C}$ need not appear in the block form (6). A parity-check matrix for $\mathcal{C}$ need not appear in the block form (7). The columns of a generator or parity-check matrix for $\mathcal{C}$, however, separate into two sets. All columns of the first set are linearly independent of all columns of the second set, and vice versa. Furthermore, if a linear combination of the columns sums to zero, the terms of this sum belonging to the first set separately sum to zero. The two sets of columns are said to be independent. (See Section 2.2 of this paper for further detail.) Since column dependences of a matrix are unaffected by premultiplication by a nonsingular matrix, we have that *a code is equivalent to a sum of two codes if and only if the columns of its* Ω*-matrices or* Λ*-matrices separate into independent sets.*

Some readily established properties of the sum just defined follow:

$$\mathcal{A} \cong \mathcal{A}' \text{ and } \mathcal{B} \cong \mathcal{B}' \quad \text{implies} \quad \mathcal{A} + \mathcal{B} \cong \mathcal{A}' + \mathcal{B}'; \tag{8}$$

$$\mathcal{A} + \mathcal{B} \cong \mathcal{B} + \mathcal{A}; \tag{9}$$

$$\mathcal{A} + (\mathcal{B} + \mathcal{C}) = (\mathcal{A} + \mathcal{B}) + \mathcal{C}; \tag{10}$$

if

$$\mathcal{C} = \mathcal{A} + \mathcal{B}, \qquad \mathcal{C}^\dagger = \mathcal{A}^\dagger + \mathcal{B}^\dagger. \tag{11}$$

1.5 *The Product of Two Codes*

We first remind the reader of the definition and elementary properties of the direct or Kronecker product of two matrices. Let $R = (r_{ij})$ be a matrix with a rows and b columns. Let $S = (s_{ij})$ be a matrix with c rows and d columns. The Kronecker product $T = R \times S$ of R times S (the order of factors is important) is the matrix of ac rows and bd columns with partitioned structure

---

* Whenever probabilities are discussed in this paper, the usual binary symmetric channel is assumed.

$$T = R \times S = \left(\begin{array}{c:c:c:c} r_{11}S & r_{12}S & \cdots & r_{1b}S \\ \hdashline r_{21}S & r_{22}S & & r_{2b}S \\ \hdashline \vdots & & & \vdots \\ \hdashline r_{a1}S & r_{a2}S & \cdots & r_{ab}S \end{array} \right).$$

The rows and columns of T can be labelled by pairs of integers so that a typical element of T is $t_{ij:kl} = r_{ik}s_{jl}$. These indexing pairs are listed in dictionary order, so that ij precedes $i'j'$ if either $i < i'$, or, when $i = i'$, if $j < j'$. For example 14 precedes 23, and 63 precedes 64.

One readily establishes the following properties for the Kronecker product:

$$Q \times (R \times S) = (Q \times R) \times S, \tag{12}$$

$$\widetilde{R \times S} = \tilde{R} \times \tilde{S}, \tag{13}$$

$$(P \times Q)(R \times S) = (PR) \times (QS), \tag{14}$$

$$R \times S = \sigma(S \times R)\mu. \tag{15}$$

In (13), the tilde indicates transpose. In (14), it is assumed that the columns of P are equinumerous with the rows of R and that the columns of Q are equinumerous with the rows of S. The product PR indicates the usual matrix product. In (15), if R has a rows and b columns and S has c rows and d columns, then σ and μ are permutation matrices of dimension ac and bd respectively and these matrices depend only on the numbers a, b, c and d and not the entries of R or S.

Let $\mathcal{A}$ be an (n,k)-code and let $\mathcal{B}$ be an (n',k')-code. We define a new code $\mathcal{C}$ by

$$\Omega(\mathcal{C}) = \Omega(\mathcal{A}) \times \Omega(\mathcal{B}). \tag{16}$$

The code $\mathcal{C}$ so defined is an (nn',kk')-code called the product of $\mathcal{A}$ and $\mathcal{B}$ and we write $\mathcal{C} = \mathcal{A}\mathcal{B}$. It is an easy consequence of the properties of the Kronecker product that $\mathcal{C}$ so defined is an (nn',kk')-code and does not depend on the particular generator matrices used for $\mathcal{A}$ and $\mathcal{B}$ in (16).

From (12) through (15) the following properties of code multiplication are readily established:

$$\mathcal{A} \cong \mathcal{A}' \text{ and } \mathcal{B} \cong \mathcal{B}' \quad \text{implies} \quad \mathcal{A}\mathcal{B} \cong \mathcal{A}'\mathcal{B}', \tag{17}$$

$$\mathcal{A}\mathcal{B} \cong \mathcal{B}\mathcal{A}, \tag{18}$$

$$\mathcal{A}(\mathcal{B}\mathcal{C}) \cong (\mathcal{A}\mathcal{B})\mathcal{C}, \tag{19}$$

$$\mathcal{A}(\mathcal{B} + \mathcal{C}) \cong \mathcal{A}\mathcal{B} + \mathcal{A}\mathcal{C}. \tag{20}$$

We note that $(\mathcal{A}\mathcal{B})^\dagger$ is not equivalent to $\mathcal{A}^\dagger\mathcal{B}^\dagger$ in general.

Let $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C} = \mathcal{A}\mathcal{B}$ be respectively an (n,k)-, an (n',k')- and an (nn',kk')-code with generator matrices Ω, Ω' and Ω'' and parity-check matrices Λ, Λ' and Λ''. There does not seem to be a simple expression for a parity-check matrix for $\mathcal{C}$ in terms of Λ and Λ'. However, if we confine our examination of codes to equivalences only, the structure of the parity checks for the product of two codes can be described simply.

We may suppose, then, that Ω and Ω' are in M-form. The structure of Ω'' is then given, up to equivalences, by

$$\begin{aligned} \Omega'' &= (I_k \,\vdots\, M) \times (I_{k'} \,\vdots\, M') \\ &\cong (I_k \times I_{k'} \,\vdots\, I_k \times M' \,\vdots\, M \times I_{k'} \,\vdots\, M \times M'). \end{aligned} \tag{21}$$

Denote the last $nn' - kk'$ columns of this last matrix by N. Then $(I_{nn'-kk'} \,\vdots\, \tilde{N})$ is the parity-check matrix for a code equivalent to $\mathcal{C}$.

It is readily seen from (21) that a code $\mathcal{C}'$ equivalent to $\mathcal{C}$ can be described as follows. The k' information places of $\mathcal{B}$ are replaced by letters (n-place binary sequences) of the code $\mathcal{A}$. This accounts for the kk' information places of $\mathcal{C}'$ and for the $k'(n - k)$ check places of $\mathcal{C}'$ described by the block $M \times I_{k'}$ in (21). The $n' - k'$ parity checks of $\mathcal{B}$ are then applied to these k' "information hyperplaces." The block $I_k \times M'$ in (21) describes repeated application of checks of $\mathcal{B}$ over the first k positions of the information hyperplaces of $\mathcal{C}'$ and accounts for $(n' - k')k$ checks. The block $M \times M'$ gives $(n - k)(n' - k')$ additional checks over the information places of $\mathcal{C}'$.

Up to equivalence, the product of two codes can be described in another, perhaps more simple, manner. Let $\mathcal{C} = \mathcal{A}\mathcal{B}$, where $\mathcal{A}$ is an (n,k)-code and $\mathcal{B}$ is an (n',k')-code. Then $\mathcal{C}$ is equivalent to the (nn',kk')-code $\mathcal{C}'$ obtained as follows. $\mathcal{A}$ is equivalent to a code $\mathcal{A}'$ with k information places and $n - k$ check places; $\mathcal{B}$ is equivalent to a code $\mathcal{B}'$ with k' information places and $n' - k'$ check places. In both $\mathcal{A}'$ and $\mathcal{B}'$, the check digits are mod 2 sums only over the information places. Write the kk' information places of $\mathcal{C}'$ in a rectangular array of k' rows and k

columns. Treat each row of the array as the k information places of a letter of $\mathcal{A}'$ and affix the corresponding check digits to obtain k' rows each of n binary digits. Regard each column of the array as the k' information places of a letter of $\mathcal{B}'$ and affix to each column the $n' - k'$ corresponding $\mathcal{B}'$ check digits. The nn' binary digits so obtained, read off in some fixed order, give the corresponding letter of $\mathcal{C}'$. It is to be noted that, in this description of $\mathcal{C}'$, $(n - k)(n' - k')$ of the check digits involve sums over other check digits, whereas in the description given by the last block of (21) these check digits are given as linear sums over the information places only.

1.6 *Arithmetic of Equivalence Classes*

The sum and product of group codes introduced in the preceding two sections provide an arithmetic of equivalence classes of codes. As before, let $\hat{\mathcal{A}}$ denote the equivalence class of codes to which the (n,k)-code $\mathcal{A}$ belongs. We define the sum of two equivalence classes by

$$\hat{\mathcal{A}} + \hat{\mathcal{B}} \equiv \widehat{(\mathcal{A} + \mathcal{B})}.$$

The self-consistency of this definition follows from (8). Similarly we define a product

$$\hat{\mathcal{A}}\hat{\mathcal{B}} \equiv \widehat{\mathcal{A}\mathcal{B}}$$

which is seen to be consistent from (17). Equations (8) through (11) and (17) through (20) give at once

$$\begin{aligned} \hat{\mathcal{A}} + \hat{\mathcal{B}} &= \hat{\mathcal{B}} + \hat{\mathcal{A}}, \\ \hat{\mathcal{A}} + (\hat{\mathcal{B}} + \hat{\mathcal{C}}) &= (\hat{\mathcal{A}} + \hat{\mathcal{B}}) + \hat{\mathcal{C}}, \\ \hat{\mathcal{A}}\hat{\mathcal{B}} &= \hat{\mathcal{B}}\hat{\mathcal{A}}, \\ \hat{\mathcal{A}}(\hat{\mathcal{B}}\hat{\mathcal{C}}) &= (\hat{\mathcal{A}}\hat{\mathcal{B}})\hat{\mathcal{C}}, \\ \hat{\mathcal{A}}(\hat{\mathcal{B}} + \hat{\mathcal{C}}) &= \hat{\mathcal{A}}\hat{\mathcal{B}} + \hat{\mathcal{A}}\hat{\mathcal{C}}. \end{aligned}$$

The simple two-letter code, **1**, consisting of the letters 0 and 1 with parameters $n = 1$, $k = 1$ and generator matrix $\Omega = (1)$ has the property

$$\mathbf{1}\hat{\mathcal{A}} = \hat{\mathcal{A}}\mathbf{1} = \hat{\mathcal{A}},$$

for all equivalence classes $\hat{\mathcal{A}}$.

1.7 *Indecomposable Codes*

To avoid repeated cumbersome statements about trivial cases, *in this section and the next we exclude from consideration codes whose generator* *matrices contain columns of zeros.* Such columns correspond to wasted digits in the code. A new code with smaller n value and the same k value can be obtained by deleting such all-zero columns. This property of possessing no columns of zeros is maintained under equivalence. If $\mathcal{A}$ possesses the property, it is not necessarily true, however, that $\mathcal{A}^\dagger$ has no columns of zeros.

It may happen that an (n,k)-code $\mathcal{A}$ is equivalent to the sum of two or more codes. In this case, we call $\mathcal{A}$ *decomposable.* If $\mathcal{A}$ is not equivalent to the sum of two or more codes, we call $\mathcal{A}$ *indecomposable.*

If $\mathcal{A}$ is decomposable, all codes equivalent to $\mathcal{A}$ are also decomposable; if $\mathcal{A}$ is indecomposable, all codes equivalent to $\mathcal{A}$ are also indecomposable. We can therefore speak of an equivalence class $\hat{\mathcal{A}}$ of codes as being either decomposable or indecomposable according as its members are or are not decomposable.

Theorem 2: Every (n,k)-code $\mathcal{A}$ is equivalent to a sum of indecomposable codes: $\mathcal{A} \cong \mathcal{A}_1 + \mathcal{A}_2 + \cdots + \mathcal{A}_m$, where $\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_m$ are indecomposable. Furthermore, this decomposition is unique in the following sense. If also $\mathcal{A} \cong \mathcal{A}_1' + \mathcal{A}_2' + \cdots + \mathcal{A}_{m'}'$, where $\mathcal{A}_1', \mathcal{A}_2', \cdots, \mathcal{A}_{m'}'$ are indecomposable, then $m = m'$, $\mathcal{A}_1 \cong \mathcal{A}_{i_1}'$, $\mathcal{A}_2 \cong \mathcal{A}_{i_2}'$, $\cdots$, $\mathcal{A}_m \cong \mathcal{A}_{i_m}'$, where $i_1, i_2, \cdots, i_m$ are the integers 1, 2, $\cdots$, m in some order.

Theorem 2 can be stated in terms of equivalence classes as follows: *Every equivalence class $\hat{\mathcal{A}}$ of codes can be expressed as a sum of indecomposable equivalence classes $\hat{\mathcal{A}} = \hat{\mathcal{A}}_1 + \hat{\mathcal{A}}_2 + \cdots + \hat{\mathcal{A}}_m$. The indecomposable summands $\hat{\mathcal{A}}_1, \hat{\mathcal{A}}_2, \cdots, \hat{\mathcal{A}}_m$ are uniquely determined apart from order by $\hat{\mathcal{A}}$.*

A further consequence of Theorem 2 is

Theorem 3 (cancellation law of addition): Let $\hat{\mathcal{A}}$, $\hat{\mathcal{B}}$ and $\hat{\mathcal{C}}$ be any three equivalence classes of group codes. Then, if $\hat{\mathcal{A}} + \hat{\mathcal{B}} = \hat{\mathcal{A}} + \hat{\mathcal{C}}$, it follows that $\hat{\mathcal{B}} = \hat{\mathcal{C}}$. (This theorem holds also when codes with columns of zeros are allowed.)

1.8 *Optimal Properties of Indecomposable Codes*

A useful property of indecomposable codes is stated in the following theorem.

Theorem 4: Let $\mathcal{A}$ be a decomposable (n,k)-code, $k < n$, with probability of no error $Q_1(\mathcal{A})$. There exists an indecomposable (n,k)-code, $\mathcal{P}$, whose probability of no error $Q_1(\mathcal{P})$ satisfies $Q_1(\mathcal{P}) \geqq Q_1(\mathcal{A})$.

In this theorem, $Q_1(\mathcal{A})$ is the probability that a letter of $\mathcal{A}$ be decoded correctly when a maximum likelihood detector is used as the decoder (see Section 1.6, Ref. 1). A similar meaning holds for $Q_1(\mathcal{P})$. The

TABLE I — VALUES

n		k									
		1		2		3		4		5	
	X =	S	R	S	R	S	R	S	R	S	R
1	$\underline{X}$ $\underline{\bar{X}}$	1 1	1 1								
2	$\underline{X}$ $\underline{\bar{X}}$	1	1	1 1							
3	$\underline{X}$ $\underline{\bar{X}}$	1	1	2 1	1 1	1 1					
4	$\underline{X}$ $\underline{\bar{X}}$	1	1	3	1	3 2	1 1	1 1			
5	$\underline{X}$ $\underline{\bar{X}}$	1	1	4	2	6 1	2 1	4 3	1 1	1 1	
6	$\underline{X}$ $\underline{\bar{X}}$	1	1	6	3	12 1	5 1	11 4	3 2	5 4	1 1
7	$\underline{X}$ $\underline{\bar{X}}$	1	1	7	4	21 1	10 1	27 5	10 4	17 8	4 3
8	$\underline{X}$ $\underline{\bar{X}}$	1	1	9	5	34	18	63 6	28 5	54 15	18 9
9	$\underline{X}$ $\underline{\bar{X}}$	1	1	11	7	54	31	134 5	71 5	163 29	71 22
10	$\underline{X}$ $\underline{\bar{X}}$	1	1	13	8	82	51	276 4	164 4	465 46	250 40
11	$\underline{X}$ $\underline{\bar{X}}$	1	1	15	10	120	79	544 3	361 3	1283 64	809 60
12	$\underline{X}$ $\underline{\bar{X}}$	1	1	18	12	174	121	1048 2	751 2	3480 89	2484 86
13	$\underline{X}$ $\underline{\bar{X}}$	1	1	20	14	244	177	1956 1	1503 1	9256 112	7240 110
14	$\underline{X}$ $\underline{\bar{X}}$	1	1	23	16	337	254	3577 1	2887 1	24282 128	20341 127
15	$\underline{X}$ $\underline{\bar{X}}$	1	1	26	19	453	356	6395 1	5393 1	62812 144	55322 143
16	$\underline{X}$ $\underline{\bar{X}}$	1	1	29	21	613	490	11217	9763	160106 145	146237 144
17	$\underline{X}$ $\underline{\bar{X}}$	1	1	32	24	808	661	19307	17273	401824 129	376725 129
18	$\underline{X}$ $\underline{\bar{X}}$	1	1	36	27	1056	882	32685	29839	992033 113	947555 113
19	$\underline{X}$ $\underline{\bar{X}}$	1	1	39	30	1361	1157	54413	50557	2.40633 91	2.32900 91

OF S_{nk}, $\bar{S}_{nk}$, R_{nk}, $\bar{R}_{nk}$

k							
6		7		8		9	
S	R	S	R	S	R	S	R
1 1							
6 5	1 1	1 1					
25 14	5 4	7 6	1 1	1 1			
99 38	31 19	35 22	7 6	8 7	1 1	1 1	
385 105	164 70	170 80	51 35	47 32	8 7	9 8	1 1
1472 273	809 220	847 312	361 190	277 151	79 59	61 44	10 9
5676 700	3749 629	4408 1285	2484 977	1775 821	751 465	436 266	121 96
22101 1794	16749 1700	24297 5632	16749 4875	12616 5098	7240 3689	3557 1948	1503 1041
87404 4579	72783 4463	143270 26792	113662 24920	102445 37191	72783 31227	34942 17934	20341 12476
350097 11635	311233 11505	901491 137493	784390 132811	957357 320663	784390 293070	428260 213773	311233 175114
1.41325 29091	1.31126 28946	5.98528 745413	5.51748 733654	10.1746 3.18608	9.09877 3.04662	6.59254 3.27631	5.51748 2.94948
5.70816 70600	5.44572 70454	41.1752 4.14506	39.2920 4.11584	119.235 34.7994	112.170 34.0492	123.425 61.2716	112.170 58.0573
22.9032 164705	22.2371 164575	287.813 22.9827	280.215 22.9120	1482.30 397.232	1434.04 393.075	2647.03 1296.46	2516.51 1261.52
90.6994 366089	89.0390 365976	2009.86 124.432	1979.34 124.268	18884.5 4558.66	18548.3 4535.64	76284.2 29032.1	59541.8 28634.1

theorem thus states that the search for best codes can be restricted to indecomposable codes when "best" means large values of Q.

Another criterion frequently used to evaluate codes is the nearest neighbor distance, d. This quantity is the smallest nonzero weight of the letters of the code. If $d = 2e + 1$, then the code can correct all combinations of e or fewer digit errors in any transmitted letter. For a given n and k, it is not necessarily true that the code with largest d value has the largest Q_1 value.

The search for codes of largest nearest neighbor distance can also be limited to indecomposable codes as a result of

Theorem 5: Let $\mathcal{A}$ *be an* (n,k)*-code,* $k < n$*, with nearest neighbor distance* $d(\mathcal{A})$*. There exists an indecomposable* (n,k)*-code,* $\mathcal{P}$*, with nearest neighbor distance* $d(\mathcal{P}) \geqq d(\mathcal{A})$.

A convenient test exists for determining whether a given Ω-matrix in M-form is the generator matrix of an indecomposable code. Two elements, m_{rs} and m_{tu}, of M are said to be *connected* if they both have value 1 and lie either in the same column or the same row of M. A *path* in M is a sequence of elements of M each of which is connected to its successor except for the last element of the sequence. In terms of these definitions, we have the following

Test: Let $\mathcal{A}$ *be an* (n,k)*-code with* $k < n$*. Then* $\mathcal{A}$ *is decomposable if and only if* M *contains a path containing elements from every row of* M.

The above test is meaningless for (n,n)-codes. The $(1,1)$-code is indecomposable. For $n \neq 1$, the (n,n)-code is decomposable.

It is easy to show from this test for decomposability that $\mathcal{A}$ is an indecomposable (n,k)-code with no column of zeros if and only if $\mathcal{A}^\dagger$ is indecomposable and has no column of zeros.

The test for decomposability can also be used to establish that $\mathcal{C} = \mathcal{A}\mathcal{B}$ is indecomposable if and only if $\mathcal{A}$ and $\mathcal{B}$ are indecomposable.

1.9 *Enumeration of Equivalence Classes*

Although we have not succeeded in parametrizing the equivalence classes of (n,k)-codes, we can systematically enumerate these classes by a modified Polya scheme.[3] The details of the method are given in Section 2.8. Here we present the results of a computation.

We shall denote by S_{nk} the number of equivalence classes of (n,k)-codes with no columns of zero.

A generator matrix for an (n,k)-code may or may not have repeated columns. The multiplicities of columns in an Ω-matrix are preserved under equivalence. Of interest are the (n,k)-codes whose Ω-matrices have no repeated columns. We denote by $\bar{S}_{nk}$ the number of equivalence classes of (n,k)-codes having no repeated columns and no columns of zeros.

We adopt an analogous notation for the number of indecomposable equivalence classes. The number of equivalence classes of indecomposable (n,k)-codes with no columns of zeros is denoted by R_{nk}. The number of equivalence classes of indecomposable (n,k)-codes with no repeated columns and no columns of zeros is denoted by $\bar{R}_{nk}$.

Table I lists values of S_{nk}, $\bar{S}_{nk}$, R_{nk} and $\bar{R}_{nk}$. The box in row n and column k contains S_{nk} in the upper left corner, $\bar{S}_{nk}$ in the lower left corner, $\bar{R}_{nk}$ in the upper right corner and $\bar{R}_{nk}$ in the lower right corner. All entries are given to six significant figures. Numbers containing a decimal point are to be multiplied by 10^6.

From a table of values of S_{nk}, one can easily construct a table of values of W_{nk}, the number of equivalence classes of (n,k)-codes (zero columns and repetition allowed). Table II is a short table of values of

TABLE II — VALUES OF N_{nk} AND W_{nk}

n		k					
		0	1	2	3	4	5
1	N	1	1				
	W	1	1				
2	N	1	3	1			
	W	1	2	1			
3	N	1	7	7	1		
	W	1	3	3	1		
4	N	1	15	35	15	1	
	W	1	4	6	4	1	
5	N	1	31	155	155	31	1
	W	1	5	10	10	5	1
6	N	1	63	651	1395	651	63
	W	1	6	16	22	16	6
7	N	1	127	2667	11811	11811	2667
	W	1	7	23	43	43	23
8	N	1	255	10795	97155	200787	97155
	W	1	8	32	77	106	77
9	N	1	511	43435	788035	3309747	3309747
	W	1	9	43	131	240	240
10	N	1	1023	174251	6347715	53743987	109221651
	W	1	10	56	213	516	705

W_{nk} along with values of N_{nk}, the total number of distinct (n,k)-codes. One has $N_{nk} = N_{nl}$, $W_{nk} = W_{nl}$, $l = n - k$. The familiar appearance of the first five rows of the W_{nk} table provides a good example of the perils of too hasty extrapolation in mathematics.

Part II — ADDITIONAL THEORY AND PROOFS OF THEOREMS OF PART I

2.1 *Proof of Theorem 1*

Theorem 1 asserts that a necessary and sufficient condition for two $k \times n$ Ω-matrices, say Ω and Ω', to be equivalent is that their columns can be placed into a one-to-one correspondence that preserves mod 2 addition of the columns.

The necessity of the condition follows trivially from the fact that equivalence means $g\Omega\sigma = \Omega'$ for some nonsingular g and some permutation matrix σ. For the one-to-one correspondence of the theorem, associate the ith column of $\Omega\sigma$, say c_i, with the ith column of Ω', say c_i', $i = 1, 2, \cdots, n$. Then $gc_i = c_i'$, $i = 1, 2, \cdots, n$. Thus, if $c_i + c_j = c_k$, then $gc_i + gc_j = gc_k$, or $c_i' + c_j' = c_k'$. Since g is nonsingular, it also follows that $c_i' + c_j' = c_k'$ implies $c_i + c_j = c_k$.

To prove the sufficiency of the condition, suppose that the columns of Ω and Ω' can be placed into a one-to-one correspondence that preserves mod 2 addition of columns. Let σ permute the columns of Ω so that the ith column of $\Omega\sigma$ corresponds to the ith column of Ω', $i = 1, 2, \cdots, n$. Let $g \in G_k$ and μ, an $n \times n$ permutation matrix, reduce $\Omega\sigma$ to M-form. Then mod 2 addition of columns is preserved between $g\Omega\sigma\mu$ and $g\Omega'\mu$ when the ith column of the former is associated with the ith column of the latter, $i = 1, 2, \cdots, n$. The first k columns of $g\Omega\sigma\mu$ are independent since the first k columns of $g\Omega\sigma'\mu$ are. Therefore the matrix g_1 formed by the first k rows and k columns of $g\Omega\sigma\mu$ is nonsingular. The matrix $g_1^{-1}g\Omega\sigma\mu$ is in M-form and, when its ith column is associated with the ith column of $g\Omega'\mu$, mod 2 addition of columns is still preserved. But then columns $k + 1, k + 2, \cdots, n$ of these two matrices are identical linear combinations of their identical first k columns, so that $g_1^{-1}g\Omega\sigma\mu = g\Omega'\mu$. It follows then that $\Omega' = g^{-1}g_1^{-1}g\Omega\sigma$, so that Ω' and Ω are equivalent.

2.2 *Decomposition of Sets of Vectors*

In this section we present five lemmas and a theorem concerning linear dependence of vectors. This material is preparatory for the proof of Theorem 2. While it is true that Theorem 2 can be proved much more directly (and abstractly) than is done here, it is felt that the procedure to be followed gives more insight into the nature of the problem at hand than do the shorter more abstract proofs.

Here we shall consider collections of vectors drawn with *possible repetitions* from a finite dimensional vector space over a finite field of scalars. In the application to be made later, the vectors will be columns taken from the generator matrix of a code, and the scalars will as usual be zero or one. The reader may, if he wishes, restrict his considerations to vectors and scalars of this sort. Throughout this section, we agree to exclude the null- or zero-vector from consideration as a member of any of the collections of vectors we may discuss.

Let $S_1, S_2, \cdots, S_m$ be nonempty finite sets of vectors. Denote the vectors of S_i by $\mathbf{v}_{ij}$, $j = 1, 2, \cdots, r_i$, for $i = 1, 2, \cdots, m$. The sets $S_1, S_2, \cdots, S_m$ are then called *independent* if every relation of the form

$$\sum_{i=1}^{m} \sum_{j=1}^{r_i} \alpha_{ij} \mathbf{v}_{ij} = 0$$

implies

$$\sum_{j=1}^{r_i} \alpha_{ij} \mathbf{v}_{ij} = 0, \qquad i = 1, 2, \cdots, m.$$

Clearly, no vector in any one such set can be written as a linear combination of vectors taken only from the other sets. Directly from the definition of independence we also have

Lemma 1: Let the sets S_i be independent and let R_i be a subset of S_i, $i = 1, 2, \cdots, m$. Then the nonempty sets among $R_1, R_2, \cdots, R_m$ are independent.

A set, S, of vectors is called *indecomposable* if S cannot be written as a union of two or more independent subsets of S. Every vector in an indecomposable set containing more than one vector can be written as a linear combination of other vectors in the set. Clearly, a set S that is not indecomposable is the union of independent indecomposable subsets, $S_1, S_2, \cdots, S_m$. In this case we say that S can be *decomposed* into independent indecomposable *components* $S_1, S_2, \cdots, S_m$.

A linear form $l = \alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \cdots + \alpha_j \mathbf{v}_j$ is called *irreducible* if no collection of $j - 1$ or fewer of the terms $\alpha_1 \mathbf{v}_1, \alpha_2 \mathbf{v}_2, \cdots, \alpha_j \mathbf{v}_j$ sums to zero; otherwise, the linear form is called *reducible*. Two linear forms are called *disjoint* if the respective sets of vectors with nonzero coefficients in the two forms are disjoint. We have then

Lemma 2: Every reducible linear form that is equal to zero is the sum of disjoint irreducible linear forms each of which is zero.

Proof: Suppose $l = \alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \cdots + \alpha_j \mathbf{v}_j$ to be reducible where all the α's are different from zero. Then there are subsets of terms of l that add to zero. Choose such a subset containing a minimal number of terms and call the sum of these terms the linear form l_1. The form l_1 must be irreducible or it would not contain a minimal number of terms. Repeat this procedure for $l - l_1 \equiv l_2 = 0$. After a finite number of steps we obtain an irreducible form l_i and $l = l_1 + l_2 + \cdots + l_i$. The forms so obtained are disjoint by construction.

Let S contain r vectors. One can form $p^r - 1$ linear forms

$$\sum_1^r \alpha_i \mathbf{v}_i$$

of these vectors where not all the α's are zero. Here p is the number of elements in the field of scalars ($p = 2$ in the applications to follow). From this list of linear forms, delete those that do not sum to zero. From the remaining forms, delete those that are reducible. One arrives then at a uniquely determined set $\mathfrak{L}$ of irreducible sums, each one of which is zero. Two vectors of S, say $\mathbf{v}_1$ and $\mathbf{v}_2$, are said to be *directly connected* to each other if they appear together as terms in any one of the irreducible sums of $\mathfrak{L}$. A vector of S not appearing in any of the linear forms of $\mathfrak{L}$ is said to be directly connected to itself. Two vectors of S, $\mathbf{v}_1$ and $\mathbf{v}_2$, are said to be *connected* if there exist vectors

$$\mathbf{v}_{i_1}, \mathbf{v}_{i_2}, \cdots, \mathbf{v}_{i_q}$$

of S such that $\mathbf{v}_1$ is directly connected to $\mathbf{v}_{i_1}$, $\mathbf{v}_{i_q}$ is directly connected to $\mathbf{v}_2$ and $\mathbf{v}_{i_\alpha}$ is directly connected to $\mathbf{v}_{i_{\alpha+1}}$, $\alpha = 1, 2, \cdots, q - 1$. If $\mathbf{v}_1$ is connected to $\mathbf{v}_2$, we write $\mathbf{v}_1 \sim \mathbf{v}_2$. Evidently, for all vectors $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ of S we have: (a) $\mathbf{v}_1 \sim \mathbf{v}_1$; (b) $\mathbf{v}_1 \sim \mathbf{v}_2$ implies $\mathbf{v}_2 \sim \mathbf{v}_1$; (c) if $\mathbf{v}_1 \sim \mathbf{v}_2$ and $\mathbf{v}_2 \sim \mathbf{v}_3$, then $\mathbf{v}_1 \sim \mathbf{v}_3$. The vectors of S are therefore uniquely separated into disjoint equivalence classes by the connectedness relation $\sim$.

Lemma 3: The totality of vectors of S belonging to an equivalence class E of connected vectors forms an indecomposable set.

For, suppose E could be written as the union of two independent subsets S_1 and S_2 of E. Since all elements of E are connected, there must be a $\mathbf{v}_1$ in S_1 and a $\mathbf{v}_2$ in S_2 such that $\mathbf{v}_1$ is directly connected to $\mathbf{v}_2$. There is therefore a linear form in $\mathfrak{L}$ of the form

$$\alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \sum_3^t \alpha_i \mathbf{v}_i = 0$$

with $\alpha_1 \neq 0$, $\alpha_2 \neq 0$. By the definition of independence, the terms in this sum belonging to S_1 add to zero, as do the terms belonging to S_2. But this contradicts the irreducibility of sums in $\mathfrak{L}$.

Lemma 4: Distinct equivalence classes $S_1, S_2, \cdots, S_m$ *of connected vectors of* S *are independent sets of vectors.*

Proof: Consider any linear form

$$l = \sum \sum \alpha_{ij} \mathbf{v}_{ij}$$

of vectors of S that is zero. Suppose l contains vectors from different equivalence classes with nonzero coefficients. Then, since $l = 0$, l cannot be irreducible, for in this case the vectors in different equivalence classes would be directly connected. Since it is reducible, l can be written by Lemma 2 as the sum of disjoint irreducible forms each of which is zero. But none of these forms can contain vectors from different equivalence classes. Adding together all the irreducible forms containing vectors from any one equivalence class, we get

$$\sum_j \alpha_{ij} \mathbf{v}_{ij} = 0, \qquad i = 1, 2, \cdots, m.$$

Lemma 5: All vectors of an indecomposable subset P *of* S *belong to the same equivalence class of connected vectors.*

For, let R_i be the set of vectors of P that belongs to the equivalence class S_i, $i = 1, 2, \cdots, m$. By Lemmas 1 and 4, the sets R_i are independent and the assumed indecomposable set P is then exhibited as the union of independent subsets. This is a contradiction unless all the R_i but one are empty.

The preceding lemmas and definitions allow us to state finally the following

Theorem 6: A set S *of vectors can be decomposed into independent indecomposable components in only one way.*

Proof: We have seen that S can be separated into equivalence classes of connected vectors in a unique manner. Lemmas 3 and 4 show these equivalence classes to be a decomposition of S into independent indecomposable sets. Suppose now that S could be decomposed in another manner into independent indecomposable sets. Lemma 5 shows that each such indecomposable set is completely contained in an equivalence class. There cannot be more than one such indecomposable set in any equivalence class, for then the equivalence class would be the union of two or more independent subsets which contradicts Lemma 3.

We point out once again in closing this section that the vectors of the set S here considered need not be distinct. S may contain several copies of a single vector of the linear vector space under consideration.

2.3 *Proof of Theorem 2*

Let us regard the columns of a generator matrix $\Omega(\mathcal{A})$ as a collection of vectors. The linear relations satisfied by a set of vectors determine whether or not the set is indecomposable. The linear relations satisfied by the column vectors of generator matrices of equivalent codes are identical (except for possible renumbering of the columns). It follows immediately that a code $\mathcal{A}$ is indecomposable if and only if the columns of any (and hence every) generator matrix $\Omega(\mathcal{A})$ form an indecomposable set of vectors. With this remark, we proceed to the proof of Theorem 2.

That every (n,k)-code $\mathcal{A}$ is equivalent to a sum of indecomposable codes follows readily from the definitions of indecomposable codes and equivalence. Here we show only that if $\mathcal{A} \cong \mathcal{A}_1 + \mathcal{A}_2 + \cdots + \mathcal{A}_m$ and $\mathcal{A} \cong \mathcal{A}_1' + \mathcal{A}_2' + \cdots + \mathcal{A}_{m'}'$ where the $\mathcal{A}_i$ and $\mathcal{A}_i'$ are indecomposable, then $m = m'$ and $\mathcal{A}_j \cong \mathcal{A}_{i_j}'$, $j = 1, 2, \cdots, m$, where $i_1, i_2, \cdots, i_m$ are the integers $1, 2, \cdots, m$ in some order.

If R, S, $\cdots$, are matrices of respective size $r \times r'$, $s \times s'$, $\cdots$, we denote by diag $(R, S, \cdots)$ the $(r + s + \cdots) \times (r' + s' + \cdots)$ partitioned matrix having R in its first r row and r' columns, S in rows $r + 1$ to $r + s$ and columns $r' + 1$ to $r' + s'$, etc., and zeros elsewhere. Set

$$\begin{aligned} \Omega &= \mathrm{diag}\,[\Omega(\mathcal{A}_1), \Omega(\mathcal{A}_2), \cdots, \Omega(\mathcal{A}_m)], \\ \Omega' &= \mathrm{diag}\,[\Omega(\mathcal{A}_1'), \Omega(\mathcal{A}_2'), \cdots, \Omega(\mathcal{A}_{m'}')]. \end{aligned} \tag{22}$$

Then, by hypothesis, $\Omega = g\Omega'\sigma$, where $\mathcal{A}_i$ is an indecomposable (n_i, k_i)-code, $i = 1, 2, \cdots, m$; $\mathcal{A}_j'$ is an indecomposable (n_j', k_j')-code, $j = 1, 2, \cdots, m'$; and

$$\sum_{i=1}^{m} k_i = \sum_{j=1}^{m'} k_j' = k,$$

$$\sum_{i=1}^{m} n_i = \sum_{j=1}^{m'} n_j' = n.$$

The columns of Ω decompose into independent indecomposable sets $S_1, S_2, \cdots, S_m$. Here S_1 consists of the first n_1 columns of Ω, S_2 consists of the next n_2 column of Ω, etc. The columns of $\Omega'\sigma$ satisfy linear relations identical with those satisfied by the columns of Ω since $\Omega = g\Omega'\sigma$, and hence, from Theorem 6, the first n_1 columns of $\Omega'\sigma$ are an indecomposable set S_1', the next n_2 columns of $\Omega'\sigma$ are an indecomposable set S_2', etc., and these sets are independent. But the columns of $\Omega'\sigma$ are a reordering of the columns of Ω' and the latter are exhibited as m' independent indecomposable sets in (22). Therefore, $m = m'$ and $n_{i_j}' =$

n_j, $j = 1, 2, \cdots, m$, where $i_1, i_2, \cdots, i_m$ are the integers $1, 2, \cdots, m$ listed in some order. It follows then that S_j' consists entirely of those columns of Ω' that contain $\Omega(\mathcal{A}_{i_j}')$, $j = 1, 2, \cdots, m$. We can then write $\Omega'\sigma = \mu\Omega''$, where μ is a $k \times k$ permutation matrix,

$$\Omega'' = \operatorname{diag}\,[\Omega(\mathcal{A}_{i_1}')\sigma_1, \Omega(\mathcal{A}_{i_2}')\sigma_2, \cdots, \Omega(\mathcal{A}_{i_m}')\sigma_m],$$

and σ_j is an $n_j \times n_j$ permutation matrix, $j = 1, 2, \cdots, m$. On setting $g'' = g\mu$, we have $g''\Omega'' = \Omega$.

Let T_1 be the matrix of the first n_1 columns of Ω, T_2 be the matrix of the next n_2 columns of Ω, etc. Let T_1'' be the matrix of the first n_1 columns of Ω'', T_2'' be the matrix of the next n_2 columns of Ω'', etc. Then $g''T_j'' = T_j$, $j = 1, 2, \cdots, m$. But T_j is of rank k_j and g'' is nonsingular, so that $k_{i_j}' \geqq k_j$. From $\sum k_j' = \sum k_j = k$, we find $k_{i_j}' = k_j$, $j = 1, 2, \cdots, m$.

Now partition g'' in rows according to $k_1, k_2, \cdots, k_m$ and in columns according to $n_1, n_2, \cdots, n_m$. Denote the ith diagonal submatrix of g'' by g_i. Then $g''\Omega'' = \Omega$ yields $g_j\Omega(\mathcal{A}_{i_j}')\sigma_j = \Omega(\mathcal{A}_j)$, $j = 1, 2, \cdots, m$. A comparison of ranks in these equations shows that the g_j are nonsingular. Therefore $\mathcal{A}_j \cong \mathcal{A}_{i_j}'$, $j = 1, 2, \cdots, m$, and the theorem is proved.

2.4 *The Test for Indecomposability*

We have seen that an (n,k)-code $\mathcal{A}$ is indecomposable if and only if the columns of any generator matrix $\Omega(\mathcal{A})$ are an indecomposable collection of vectors. If $\Omega(\mathcal{A})$ is in M-form its first k columns are independent and each contains a single one. The other columns of $\Omega(\mathcal{A})$ can each be expressed as an irreducible sum of these first k columns. From Section 2.2 it follows that the columns of $\Omega(\mathcal{A})$ will form an indecomposable set of vectors if and only if the first k columns of $\Omega(\mathcal{A})$ are connected to each other. The reader can readily translate this statement into the test described in Section 1.8.

2.5 *Proof of Theorem 3*

The hypothesis $\hat{\mathcal{A}} + \hat{\mathcal{B}} = \hat{\mathcal{A}} + \hat{\mathcal{C}}$ means that, for codes $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ respectively in $\hat{\mathcal{A}}$, $\hat{\mathcal{B}}$ and $\hat{\mathcal{C}}$,

$$\mathcal{A} + \mathcal{B} \cong \mathcal{A} + \mathcal{C}.$$

Then

$$\begin{aligned} &\mathcal{A}_1 + \mathcal{A}_2 + \cdots + \mathcal{A}_\alpha + \mathcal{B}_1 + \mathcal{B}_2 + \cdots + \mathcal{B}_\beta \\ &\qquad\qquad \cong \mathcal{A}_1 + \mathcal{A}_2 + \cdots + \mathcal{A}_\alpha + \mathcal{C}_1 + \mathcal{C}_2 + \cdots + \mathcal{C}_\gamma, \end{aligned}$$

where the $\mathcal{A}_j$, $\mathcal{B}_j$ and $\mathcal{C}_j$ are the (unique) indecomposable code components respectively of $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$. By Theorem 2 we have $\beta = \gamma$, and there is a one-to-one correspondence set up by the equivalence relation $\cong$ between elements of the set $H_1 = \{\mathcal{A}_1, \cdots, \mathcal{A}_\alpha, \mathcal{B}_1, \cdots, \mathcal{B}_\beta\}$ and the set $H_2 = \{\mathcal{A}_1, \cdots, \mathcal{A}_\alpha, \mathcal{C}_1, \cdots, \mathcal{C}_\beta\}$. If all the $\mathcal{B}$'s map into $\mathcal{C}$'s in this correspondence, then $\sum \mathcal{B}_i \cong \sum \mathcal{C}_i$, $\hat{\mathcal{B}} = \hat{\mathcal{C}}$, and the theorem is proved. Suppose then that $\mathcal{B}_1$ maps into $\mathcal{A}_{i_1}$ of H_2. If $\mathcal{A}_{i_1}$ of H_1 maps into a $\mathcal{C}$, say $\mathcal{C}_1$, then $\mathcal{B}_1 \cong \mathcal{A}_{i_1} \cong \mathcal{C}_1$, and we go on to examine another $\mathcal{B}$ of H_1. If, however, $\mathcal{A}_{i_1}$ of H_1 maps into $\mathcal{A}_{i_2}$ of H_2, we then consider $\mathcal{A}_{i_2}$ in H_1. Proceeding in this manner, we must ultimately reach an $\mathcal{A}$ in H_1 that is mapped onto a $\mathcal{C}$, since the $\mathcal{A}$'s in H_1 and H_2 are equinumerous and $\mathcal{B}_1$ of H_1 is mapped onto an $\mathcal{A}$ of H_2. This yields a chain of equivalences starting with $\mathcal{B}_1$ and ending with a $\mathcal{C}$. Each $\mathcal{B}$ then is equivalent to a $\mathcal{C}$ and, by reversing the argument, we find a one-to-one equivalence correspondence among the $\mathcal{B}$'s and $\mathcal{C}$'s. It follows then that $\hat{\mathcal{B}} = \hat{\mathcal{C}}$.

2.6 *Proof of Theorem 4*

Theorem 4 states that if $\mathcal{A}$ is an indecomposable (n,k)-code, $k < n$, with probability of no error $Q_1(\mathcal{A})$, then there exists an indecomposable (n,k)-code, $\mathcal{P}$, with probability of no error $Q_1(\mathcal{P}) \geqq Q_1(\mathcal{A})$.

Proof: The given code $\mathcal{A}$ is equivalent, by Theorem 2, to a code $\mathcal{A}'$ that is the sum of indecomposable codes:

$$\mathcal{A}' = \mathcal{B}_1 + \mathcal{B}_2 + \cdots + \mathcal{B}_m,$$

where $\mathcal{B}_i$ is an indecomposable (n_i, k_i)-code and $\sum k_i = k$, $\sum n_i = n$. Let $\mathcal{B}_i$ have probability of no error $Q_1(\mathcal{B}_i)$ when used with a maximum likelihood detector. Then $\mathcal{A}'$ has probability of no error $Q_1(\mathcal{A}') = Q_1(\mathcal{B}_1)Q_1(\mathcal{B}_2) \cdots Q_1(\mathcal{B}_m)$. [See remark following (7).]

We shall show below that the theorem is true for $m = 2$. The proof for general m then follows readily by induction. For, suppose the theorem to be true for $m = 2, 3, \cdots, r$. If then $\mathcal{A}' = \mathcal{B}_1 + \mathcal{B}_2 + \cdots + \mathcal{B}_r + \mathcal{B}_{r+1}$, by the induction hypothesis there is an indecomposable $(n - n_{r+1}, k - k_{r+1})$-code $\mathcal{B}'$ with $Q_1(\mathcal{B}') \geqq Q_1(\mathcal{B}_1)Q_1(\mathcal{B}_2) \cdots Q_1(\mathcal{B}_r)$. The decomposable code $\mathcal{A}'' = \mathcal{B}' + \mathcal{B}_{r+1}$ has probability of no error $Q_1(\mathcal{A}'') = Q_1(\mathcal{B})Q_1(\mathcal{B}_{r+1})$. Again by the induction hypothesis, there exists an indecomposable (n,k)-code, $\mathcal{P}$, with $Q_1(\mathcal{P}) \geqq Q_1(\mathcal{A}'') = Q_1(\mathcal{B}')Q_1(\mathcal{B}_{r+1}) \geqq Q_1(\mathcal{B}_1)Q_1(\mathcal{B}_2) \cdots Q_1(\mathcal{B}_r)Q_1(\mathcal{B}_{r+1}) = Q_1(\mathcal{A}')$. The theorem is then true also for $m = 2, 3, \cdots, r + 1$.

To prove the theorem for $m = 2$, we distinguish two cases. First suppose $n_2 \neq 1$. We can suppose the generator matrices for $\mathcal{B}_1$ and $\mathcal{B}_2$ written in M-form so that a generator matrix for $\mathcal{A}'$ has the form

$$\Omega(\mathcal{A}') = \left(\begin{array}{c|c|c|c} I_{k_1} & M_1 & 0 & 0 \\ \hline 0 & 0 & I_{k_2} & M_2 \end{array}\right). \tag{23}$$

Consider now the (n,k)-code $\mathcal{P}$ with generator matrix

$$\Omega(\mathcal{P}) = \left(\begin{array}{c|c|c|c} & & & 11\cdots1 \\ & & & 00\cdots0 \\ I_{k_1} & M_1 & 0 & \vdots \quad \vdots \\ & & & 00\cdots0 \\ \hline 0 & 0 & I_{k_2} & M_2 \end{array}\right), \tag{24}$$

where the upper right section of $\Omega(\mathcal{P})$ has one row of 1's and $k_1 - 1$ rows of zeros. We observe first that $\mathcal{P}$ is indecomposable, since $\mathcal{P}$ is equivalent to a code with generator matrix in M-form with

$$M = \left(\begin{array}{c|c} & 11\cdots1 \\ & 00\cdots0 \\ M_1 & \vdots \quad \vdots \\ & 00\cdots0 \\ \hline 0 & M_2 \end{array}\right).$$

Since $\mathcal{B}_1$ and $\mathcal{B}_2$ are indecomposable, both M_1 and M_2 have paths that contain all their rows, by the test of Section 1.8. A single path containing all rows of M is then easily obtained by joining together the paths for M_1 and M_2 by some of the ones of the upper right block of M. The code associated with M is thus indecomposable, and so is $\mathcal{P}$.

The last $k_1 - 1$ rows of $\Omega(\mathcal{B}_1)$ generate an $(n_1, k_1 - 1)$-code. Let the letters of this code be B_{11}', B_{12}', $\cdots$, $B_{1\sigma}'$, where $\sigma = 2^{k_1-1}$. Let the first row of $\Omega(\mathcal{B}_1)$ be denoted by B_{11}. Then the $\mu_1 = 2^{k_1}$ letters of $\mathcal{B}_1$ are B_{11}', B_{12}', $\cdots$, $B_{1\sigma}'$ and $B_{11} + B_{11}'$, $B_{11} + B_{12}'$, $\cdots$, $B_{11} + B_{1\sigma}'$. Let the letters of $\mathcal{B}_2$ be B_{21}, B_{22}, $\cdots$, $B_{2\mu_2}$ where $\mu_2 = 2^{k_2}$. Then the letters of $\mathcal{A}'$ can be denoted by the $\mu_1\mu_2$ symbols (B_{1i}', B_{2j}) and $(B_{11} + B_{1i}', B_{2j})$, where $i = 1, 2, \cdots, \sigma$ and $j = 1, 2, \cdots, \mu_2$. The notation here is that (B_{1i}', B_{2j}) stands for the sequence B_{1i}' followed by the sequence B_{2j}, for example.

In the notation just introduced, the $\mu_1\mu_2$ letters of $\mathcal{P}$ are (B_{1i}, B_{2j}) and $(B_{11} + B_{1i}', \bar{B}_{2j})$, where $i = 1, 2, \cdots, \sigma$ and $j = 1, 2, \cdots, \mu_2$ and $\bar{B}_{2j}$ denotes the sequence B_{2j} with its last $n_2 - k_2 = l_2$ places complemented.

That is, $\bar{B}_{2j}$ is obtained from B_{2j} by changing to zero every one in the last l_2 places of B_{2j} and by changing to one every zero in the last l_2 places of B_{2j}.

Consider now transmitting with $\mathcal{P}$ over a binary symmetric channel using the following decoding rules. Apply the maximum likelihood detector for $\mathcal{B}_1$ to the first n_1 digits of a received sequence R. One thus obtains a letter of $\mathcal{B}_1$, say B_{1i}. If B_{1i} is one of the letters B_{11}', B_{12}', $\cdots$, $B_{1\sigma}'$, apply the maximum likelihood detector for $\mathcal{B}_2$ to the last n_2 places of R to obtain a letter of $\mathcal{B}_2$, say B_{2j}. The pair (B_{1i}, B_{2j}) is taken as the decoded version of R. If,however, B_{1i} is one of the letters $B_{11} + B_{11}'$, $B_{11} + B_{12}'$, $\cdots$, $B_{11} + B_{1\sigma}'$, complement the last l_2 places of R, and then apply the maximum likelihood detector of $\mathcal{B}_2$ to the last n_2 digits of this new sequence derived from R. A letter B_{2j}, say, of $\mathcal{B}_2$ will be obtained. The decoded version of R is taken to $(B_{1i}, \bar{B}_{2j})$.

It is readily seen that on using the indecomposable code $\mathcal{P}$ with this decoding scheme, the probability of no error is $Q_1(\mathcal{B}_1)Q_1(\mathcal{B}_2)$. Since the maximum likelihood detector for $\mathcal{P}$ must do as well, $Q_1(\mathcal{P}) \geqq Q_1(\mathcal{B}_1) \cdot Q_1(\mathcal{B}_2) = Q_1(\mathcal{A}') = Q_1(\mathcal{A})$, and the theorem is proved for this case.

If $n_2 = 1$, but $n_1 \neq 1$, reverse the roles of $\mathcal{B}_1$ and $\mathcal{B}_2$ in the preceding argument. The case $n_1 = n_2 = 1$ has been excluded by the condition $k < n$, for $n_1 = n_2 = 1$ implies $k_1 = k_2 = 1$, or $n = k = 2$.

This completes the proof.

2.7 *Proof of Theorem 5*

The nearest neighbor distance, $d(\mathcal{A})$, of a group code $\mathcal{A}$ is the smallest of the nonzero weights of the letters of $\mathcal{A}$. If $\mathcal{A}$ and $\mathcal{A}'$ are equivalent, $d(\mathcal{A}) = d(\mathcal{A}')$, and indeed the list of weights of letters of $\mathcal{A}$ is the same set of numbers as the list of weights of the letters of $\mathcal{A}'$. It is easy to see that if $\mathcal{A} = \mathcal{B} + \mathcal{C}$ then $d(\mathcal{A}) = \min\,[d(\mathcal{B}), d(\mathcal{C})]$. Thus, if $\mathcal{A} \cong \mathcal{B}_1 + \mathcal{B}_2 + \cdots + \mathcal{B}_m$, $\mathrm{d}(\mathcal{A}) = \min\,[d(\mathcal{B}_1), d(\mathcal{B}_2), \cdots, d(\mathcal{B}_m)]$.

The proof of Theorem 5 follows the outline of the proof of Theorem 4. The inductive part of the proof only requires substituting d's for Q's. The pertinent equations are:

$$\begin{aligned} d(\mathcal{B}') &\geqq \min\,[d(\mathcal{B}_1), d(\mathcal{B}_2), \cdots, d(\mathcal{B}_r)], \\ d(\mathcal{A}'') &= \min\,[d(\mathcal{B}'), d(\mathcal{B}_{r+1})], \\ d(\mathcal{P}) \geqq d(\mathcal{A}'') &= \min\,[d(\mathcal{B}'), d(\mathcal{B}_{r+1})] \\ &\geqq \min\{\min\,[d(\mathcal{B}_1), \cdots, d(\mathcal{B}_r)], d(\mathcal{B}_{r+1})\} \\ &= \min\,[d(\mathcal{B}_1), \cdots, d(\mathcal{B}_{r+1})] = d(\mathcal{A}') = d(\mathcal{A}). \end{aligned}$$

To prove the theorem for $m = 2$, we again consider a generator matrix for $\mathcal{A}'$ in the form given by (23). Without loss of generality, we suppose $d(\mathcal{A}') = d(\mathcal{B}_1)$, so that $d(\mathcal{B}_1) \leqq d(\mathcal{B}_2)$. Now suppose $l_2 = n_2 - k_2 \geqq 1$. We compare $\mathcal{A}'$ with the indecomposable code $\mathcal{P}$ given by (24). The nonzero letters of $\mathcal{P}$ are the $2^{k_1+k_2} - 1$ nontrivial linear combinations of the rows of $\Omega(\mathcal{P})$. Every such linear combination that contains one or more of the first k_1 rows of $\Omega(\mathcal{P})$ has weight $\geqq d(\mathcal{B}_1)$, since the first n_1 places will be a nonzero letter of $\mathcal{B}_1$ and the last n_2 places have weight $\geqq 0$. Every linear combination of rows of $\Omega(\mathcal{P})$ that does not contain any of the first k_1 rows is just a letter of $\mathcal{B}_2$ preceded by n_1 zeros, and hence has weight $\geqq d(\mathcal{B}_2) \geqq d(\mathcal{B}_1)$. We thus have $d(\mathcal{P}) \geqq d(\mathcal{B}_1) = d(\mathcal{A}')$.

If $l_2 = 0$, then $k_2 = n_2 = 1$, since $\mathcal{B}_2$ is assumed indecomposable. Then $d(\mathcal{B}_2) = 1$ and, since $d(\mathcal{B}_1) \leqq d(\mathcal{B}_2)$, $d(\mathcal{A}') = d(\mathcal{B}_1) = 1$. However, for every indecomposable (n,k)-code $\mathcal{P}$, we have $d(\mathcal{P}) \geqq 1 = d(\mathcal{A}')$, and so the theorem is proved for $m = 2$.

2.8 *Enumeration Formulae*

Let G be a finite group with elements $g_1, g_2, \cdots, g_r$, where r is the order of G. Define $g_i \sim g_j$ if there exists an element $g \in G$ such that $g_i = gg_jg^{-1}$. The equivalence relation $\sim$ partitions G into equivalence classes $C_1, C_2, \cdots, C_p$ called *classes of conjugate elements.* Now suppose that corresponding to each element g_i of G there is a permutation, $\sigma(g_i)$, of m objects $S_1, S_2, \cdots, S_m$ of a set S such that if $g_ig_j = g_k$, then $\sigma(g_i)\sigma(g_j) = \sigma(g_k)$. We define two of the objects of the collection S, say S_i and S_j, to be equivalent if there is a $\sigma(g_l)$, $g_l \in G$, that replaces S_i by S_j. The collection of objects S is then partitioned into equivalence classes. A well-known theorem (p. 231, Ref. 3) gives, for the number of equivalence classes N of S,

$$N = \frac{1}{r}\sum_{i=1}^{p} n(C_i)\chi(C_i). \tag{25}$$

Here $n(C_i)$ is the number of elements of G in the equivalence class C_i and $\chi(C_i)$ is the number of elements of S left invariant by any $\sigma(g_i)$, $g_i \in C_i$. [It is easy to show that if $g_i \sim g_j$, then $\sigma(g_i)$ and $\sigma(g_j)$ leave the same number of elements of S invariant.]

We apply this theorem to the enumeration of (n,k)-codes as follows. For the group G we choose the collection G_k of nonsingular $k \times k$ matrices (mod 2) of order

$$|G_k| = (2^k - 2^0)(2^k - 2^1) \cdots (2^k - 2^{k-1}). \tag{26}$$

Let $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{2^k-1}$ be the nonzero k-place binary column vectors. For the sets $S_1, S_2, \cdots, S_m$ we choose the $m = (2^k - 1)^n$ possible collections of the $\mathbf{v}$'s taken n at a time (repetitions of $\mathbf{v}$'s within any S allowed). The elements of G_k permute the $2^k - 1$ vectors $\mathbf{v}$ among themselves by ordinary matrix multiplication. That is, if $g_i\mathbf{v}_j = \mathbf{v}_l$, we say that g_i induces a permutation $\mu(g_i)$ that replaces $\mathbf{v}_j$ by $\mathbf{v}_l$. The permutation $\mu(g_i)$ of the $\mathbf{v}$'s in turn induces a permutation $\sigma(g_i)$ of the sets $S_1, S_2, \cdots, S_m$. We note that if $n \leqq 2^k - 1$, then

$$\bar{m} = \binom{2^k - 1}{n}$$

of the m S's have the property of containing only distinct vectors (no repetitions), and these $\bar{m}$ special S's are permuted among themselves under $\sigma(g_i)$. We denote by $\bar{\sigma}(g_i)$ the permutation of these $\bar{m}$ special S's induced by g_i.

We now define two $k \times n$ binary matrices Ω and Ω', regardless of their rank, to be equivalent if there exists a $g \in G_k$ and an $n \times n$ permutation matrix ν such that $\Omega' = g\Omega\nu$. The number of equivalence classes of $k \times n$-matrices none of which has columns of zeros is then clearly the same as the number of equivalence classes of the sets $S_1, \cdots, S_m$. Applying (25), we write

$$T_{nk} = \frac{1}{|G_k|} \sum_i n(C_i)\chi(C_i), \tag{27}$$

$$\bar{T}_{nk} = \frac{1}{|G_k|} \sum_i n(C_i)\bar{\chi}(C_i), \tag{28}$$

where $|G_k|$ is given by (26), $n(C_i)$ is the number of elements of G_k in class C_i, and $\chi(C_i)$ and $\bar{\chi}(C_i)$ are the number of objects left invariant respectively by $\sigma(g_i)$ and $\bar{\sigma}(g_i)$, $g_i \in C_i$. The quantities T_{nk} and $\bar{T}_{nk}$ are, respectively, the number of equivalence classes of $k \times n$ matrices with no columns of zeros and the number of equivalence classes of $k \times n$ matrices with no columns of zeros and no repeated columns.

The matrices Ω in the above enumeration may have rank less than k. It is easy to show, however, that

$$S_{nk} = T_{n,k} - T_{n,k-1}, \tag{30}$$

$$\bar{S}_{nk} = \bar{T}_{n,k} - \bar{T}_{n,k-1}, \tag{31}$$

$k = 2, \cdots, n$, $n = 1, 2, \cdots$, where, as in Section 1.9, S_{nk} and $\bar{S}_{nk}$ are, respectively, the number of equivalence classes of (n,k)-codes with no column of zeros and the number with neither repeated columns nor columns of zeros. We also have $S_{n1} = 1$ for $n = 1, 2, \cdots$ and $\bar{S}_{11} = 1, \bar{S}_{n1} = 0$ for $n > 1$.

The group G_k has been well studied, and the detail needed to evaluate (27) and (28) can be taken from the literature. Here we omit all derivations and only present such definitions and formulae as needed for our purpose. The structure of G_k is given in detail by Dickson;[4] a recipe for getting the cycle structure of the permutations of the **v**'s induced by elements of G_k is given by Elspas.[5]

A polynomial of degree $d > 0$,

$$P(x) = x^d + a_1x^{d-1} + a_2x^{d-2} + \cdots + a_d ,$$

where the a's are zero or one, is said to be irreducible if it cannot be written as the product of two or more polynomials with coefficients zero or one, where each factor is of degree greater than zero. (All addition of coefficients is to be done mod 2.) For each d there are a finite number of irreducible polynomials. In what follows, we shall exclude from consideration the irreducible polynomial $P(x) = x$. The first few irreducible polynomials are $x + 1$, $x^2 + x + 1$, $x^3 + x + 1$, $x^3 + x^2 + 1$. A more comprehensive table of irreducible polynomials is given by Church,[6] where, for each irreducible polynomial, P, there is also listed the smallest integer e such that P divides $x^e - 1$. We suppose the irreducible polynomials to be numbered, and denote them by $P_1, P_2, P_3, \cdots$. We let d_i denote the degree of P_i and e_i denote the smallest integer e such that P_i divides $x^e - 1$. We further let t_d be the number of irreducible polynomials of degree d or less.

A partition of an integer α into positive integral parts $\lambda_1, \lambda_2, \cdots$, say $\alpha = \lambda_1 + \lambda_2 + \cdots + \lambda_p$, can also be written in the form

$$\alpha = 1\alpha_1 + 2\alpha_2 + \cdots + \alpha\alpha_\alpha = \sum_1^\alpha i\alpha_i .$$

Here α_i designates how many parts have the value i. We shall use boldface Greek letters to denote partitions. The absolute value sign will denote the value of the integer being partitioned. For example, $\boldsymbol{\alpha}$ will denote a particular partition,

$$\sum_1^\alpha i\alpha_i ,$$

of the integer $\alpha = |\boldsymbol{\alpha}|$. When dealing with many partitions $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3$, etc., we shall denote the numbers of parts of various size of $\boldsymbol{\alpha}_i$ by α_{i1}, α_{i2}, etc., so that

$$|\boldsymbol{\alpha}_i| = \sum_{j=1}^{|\boldsymbol{\alpha}_i|} j\alpha_{ij} .$$

We admit the single partition of zero, **0**, into one part. For this partition, all α's are zero.

The classes of conjugate elements of G_k can be specified conveniently by t_k-place symbols. The ith place in such a class symbol corresponds to the ith of the irreducible polynomials of degree $\leqq k$. Each place in such a class symbol is occupied by a partition. If the symbol for a class of G_k is

$$(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_{t_k}), \tag{32}$$

we require

$$\sum_{i=1}^{t_k} |\boldsymbol{\alpha}_i| d_i = k. \tag{33}$$

The various classes of G_k are given by all the distinct symbols (32) that can be formed subject to (33). The sums in (27) and (28) are over such class symbols.

We now give a recipe for the integers $n(C)$ of (27) and (28). (See p. 235, Ref. 4.) We first write

$$n(C) = \frac{|G_k|}{D(C)}.$$

Then, if C is specified by (32),

$$D(C) = \prod_{j=1}^{t_k} f(\boldsymbol{\alpha}_j, d_j).$$

Here

$$f(\boldsymbol{\alpha}_i, j) = 2^{j\theta(\boldsymbol{\alpha}_i)} \prod_{l=1}^{|\boldsymbol{\alpha}_i|} \Omega(\alpha_{il}, j),$$

where

$$\Omega(r,j) = (2^{rj} - 2^{0j})(2^{rj} - 2^{1j}) \cdots (2^{rj} - 2^{(r-1)j})$$

and

$$\theta(\boldsymbol{\alpha}_i) = \sum_{j=1}^{|\boldsymbol{\alpha}_i|} \alpha_{ij}^{\;2}(j-1) + 2\sum_{j=1}^{|\boldsymbol{\alpha}_i|-1} j\alpha_{ij} \sum_{l=j+1}^{|\boldsymbol{\alpha}_i|} \alpha_{il}.$$

To compute the quantities $\chi(C_i)$ and $\tilde{\chi}(C_i)$ of (27) and (28), we need to know the cycle structure of the permutation of the $\mathbf{v}$'s induced by an element of class C_i of G_k. Let an element of C_i, as given by (32), permute the $\mathbf{v}$'s into ν_i cycles of length i, where $i = 1, 2, \cdots, 2^k - 1$. An algorithm for finding the ν's is given by Elspas.[5] Introduce indeterminates $z_1, z_2, \cdots$, and define the product of two z's by the rule

$$z_a z_b = c z_d ,$$

where c is the greatest common divisor of a and b and d is the least common multiple of a and b. Then the ν's may be obtained from

$$z_1 + \sum_{l=1}^{2^k-1} \nu_l(C) z_l = \prod_{i=1}^{t_k} \prod_{j=1}^{|\alpha_i|} H(i,j)\alpha_{ij} ,$$

where the linear forms $H(i,j)$ in the z's are obtained recursively by

$$H(i,j) = H(i,j-1) + \frac{2^{d_i(j-1)}(2^{d_i}-1)}{q_{ij}} z_{q_{ij}},$$

$$i = 1, 2, \cdots ,$$

$$q_{ij} = e_i 2^{b_j},$$

where b_j is the smallest integer such that $2^{b_j} \geqq j$, and $H(i,0) = z_1$, $i = 1, 2, \cdots$.

An element of G_k permutes the $\mathbf{v}$'s in cycles. A collection S_j of n $\mathbf{v}$'s will remain invariant under this permutation only if S_j is composed of complete sets of the $\mathbf{v}$'s that are permuted in cycles. It is not hard to determine the number of S_j that remain fixed when the cycle structure of the permutation of the $\mathbf{v}$'s is given. We write only the final result:

$$\sum_0^\infty T_{nk} t^n = \frac{1}{|G_k|} \sum_i n(C_i) \prod_{j=1}^{2^k-1} (1 - t^j)^{-\nu_j(C_i)},$$

$$\sum_0^\infty T_{nk} t^n = \frac{1}{|G_k|} \sum_i n(C_i) \prod_{j=1}^{2^k-1} (1 + t^j)^{\nu_j(C_i)}.$$

The utterly formidable series of formulae and algorithms from (32) on were used, along with (30) and (31), to compute the S_{nk} and $\bar{S}_{nk}$ given on Table I. The R_{nk} were found from the S_{nk} by a generating function scheme which will not be described in detail here. When the R_{nk} are known for $k = 1, 2, \cdots, k_0$ and $n = 1, 2, \cdots, n_0$, these numbers can be used to find the number of equivalence classes of decomposable $(n_0 + 1, k_0)$-codes, $(n_0, k_0 + 1)$-codes and $(n_0 + 1, k_0 + 1)$-codes. By subtracting the number of decomposable equivalence classes from the appropriate S_{nk}, new values of R_{nk} are found.

The programming of these formulae for the IBM 704 presented a number of interesting problems. All quantities involved are integers. In the program, they were maintained as integers. The division indicated in (27) then provides a check as to the accuracy of the sum. Unfortunately, the integers involved are frequently enormous. Modest answers in Table I of magnitude 10^1 to 10^2 were obtained as the result of computations involving integers of magnitude 10^{30}. The total machine time needed to compute the results presented was about 45 minutes.

2.9 *An Alternate Approach to Enumeration*

In Ref. 1 we regarded any subgroup of order 2^k of the group B_n of n-place binary sequences under mod 2 addition as an (n,k)-code. Thus codes with columns of zeros were admitted. It was also pointed out that G_n is the group of automorphisms of B_n. If we regard the elements of B_n as column vectors, then multiplication of each element of B_n by an $n \times n$ matrix $g \in G_n$ sends the element into a new element of B_n and this defines the automorphism associated with g.

In an automorphism of B_n, subgroups of B_n are sent into subgroups. We denote by $g\mathcal{A}$ the subgroup into which the (n,k)-code $\mathcal{A}$ is sent under the automorphism g. As g runs through G_n, $g\mathcal{A}$ runs through all N_{nk} (n,k)-codes.

Now let H be the subgroup of G_n that leaves $\mathcal{A}$ invariant, i.e., H consists of all those elements $g \in G_n$ for which $g\mathcal{A} = \mathcal{A}$. Let S_n be the subgroup of G_n consisting of all $n!$ $n \times n$ permutation matrices. Then the elements S_nH (the collection of distinct elements of G_n obtained by multiplying every element of S_n on the right by every element of H) send $\mathcal{A}$ into an equivalent code, and it is easy to show that S_nH contains all elements of G_n that send $\mathcal{A}$ into an equivalent code. Let $g_2 \in G_n$ send $\mathcal{A}$ into a nonequivalent code $\mathcal{A}_2$. Then $g_2 \notin S_nH$. Every element of the collection S_ng_2H (i.e., all elements sg_2h with $s \in S_n$, $h \in H$) then sends $\mathcal{A}$ into a code equivalent to $\mathcal{A}_2$, and again it is easily shown that every element of G_n that sends $\mathcal{A}$ into a code equivalent to $\mathcal{A}_2$ is contained in S_ng_2H.

A collection of the form S_ngH is called a *double coset* of G_n with respect to S_n and H. Two double cosets of G_n with respect to S_n and H, say S_ng_1H and S_ng_2H, are either disjoint or identical. The group G_n can thus be decomposed into disjoint double cosets S_ng_1H, S_ng_2H, $\cdots$, S_ng_pH. The argument of the preceding paragraph can be continued to show that p, the number of double cosets of G_n with respect to S_n and H, is the number, W_{nk}, of equivalence classes of (n,k)-codes (zero columns permitted).

The following formula[7] for the number, p, of double cosets of a finite group G of order $|G|$ with respect to the subgroups H_1 and H_2 respectively of order $|H_1|$ and $|H_2|$,

$$p = \frac{|G|}{|H_1|\,|H_2|} \sum_i \frac{n_1(C_i)n_2(C_i)}{n(C_i)}, \tag{34}$$

could then be applied to the case at hand to compute W_{nk}. In (34) the sum is over the classes C_i of conjugate elements of G, $n(C_i)$ is the number of elements of G in class C_i, and $n_j(C_i)$ is the number of elements of C_i that lie in H_j, $j = 1, 2$. An appropriate choice for $\mathcal{A}$ in the enumeration in question would be the (n,k)-code whose last $n - k$ columns are zero. The set of all matrices of G_n whose last $n - k$ rows contain only zero in their first k columns then makes up the subgroup H. We do not carry out the details of the enumeration by this method further here.

2.10 *Equivalence for M-forms*

We have commented in Section 1.2 that two equivalent Ω-matrices both in M-form may have different M-matrices. It is natural to inquire into the different M-forms possible for Ω-matrices within an equivalence class.*

The M-forms of all matrices equivalent to Ω can be obtained as follows. Make any permutation of the columns of Ω that causes the resultant matrix, Ω', to have its first k columns linearly independent. Premultiply Ω' by the inverse of the matrix formed by its first k columns.

Now let

$$\Omega = \begin{pmatrix} 100\cdots0 & m_{11}m_{12}\cdots m_{1l} \\ 010\cdots0 & m_{21}m_{22}\cdots m_{2l} \\ \vdots & \vdots \\ 000\cdots1 & m_{k1}m_{k2}\cdots m_{kl} \end{pmatrix} = (I_k \vdots M),$$

where $l = n - k$. The permutations of the columns of Ω that replace its first k columns by independent columns can be generated by repeated applications of three types of elementary permutations: (a) interchange of position of two among the last l columns of Ω; (b) interchange of position of two among the first k columns of Ω; (c) interchanging one of the first k columns with one of the last l columns. A type (a) transposition is a column transposition of M and Ω is still in M-form. A type (b) transposition involving columns i and j yields a matrix that can be brought into M-form by premultiplication by the permutation matrix that interchanges rows i and j. The new M differs from the old only by interchange of rows i and j. A type (c) transposition, which interchanges column j of M with column i of I_k, is valid only if $m_{ij} = 1$ (otherwise the first k columns of the new Ω would not be independent). Let such a transposition send Ω into Ω'. Let column j of M have ones in rows i, p_1, p_2, $\cdots$, p_r and zeros elsewhere. Then Ω' can be brought into M-form by premultiplication by a matrix that adds row i of Ω' to rows $p_1, p_2, \cdots, p_r$. The new M-matrix is then obtained from the original M-matrix by these operations: leave column j unchanged; except in column j, add row i to rows $p_1, p_2, \cdots, p_r$. We call this a *pivotal operation on M about the position* m_{ij}, provided $m_{ij} = 1$.

---

* The equivalence described here has been investigated independently and in a more general setting by Tucker.[8]

Define two M-matrices to be equivalent if one can be obtained from the other by repeated applications in any order of permutations of rows or columns or by pivotal operations. Then two Ω-matrices are equivalent if and only if when reduced to M-form their M-matrices are equivalent. Equivalent M-matrices, when prefixed by a unit matrix, yield equivalent Ω-matrices. We have not been able to find a systematic method of reducing a given $k \times l$ binary matrix to a canonical form by means of pivotal operations and permutations of rows and columns.

2.11 *Miscellaneous Comments and Problems*

The Q for the sum of two codes is the product of the Q's for the summands. What is the relationship for the Q of a product in terms of the Q's of the factors? What is the relationship between the Q of a code and the Q of its dual? Answers to both of these questions probably require some detailed knowledge of the structure of the codes involved beyond a mere statement of their Q's. What detail must be known?

Decomposition of codes with respect to addition has been explored. Certain optimal properties of indecomposable codes and a unique decomposition theorem have been proved. Decomposition with respect to multiplication can be defined in a similar manner. Do analogous theorems hold in this case?

When $n < 2^k - 1$, an Ω-matrix need not have repeated columns. If an indecomposable Ω-matrix does have repeated columns, the corresponding code can be viewed as having several check digits that are identical linear combinations of the information places. Intuitively, this seems like a wasteful use of the check digits. Is it possible to prove a theorem to the effect that if $n < 2^k - 1$, there is an (n,k)-code with no repeated columns with a Q as great as that for any (n,k)-code with repeated columns? All cases of known best group codes with $n < 2^k - 1$ have no repeated columns.

A strong statement about group codes with no repeated columns that might be conjectured is the following: "Let $\mathcal{A}$ be an (n,k)-code with $n < 2^k - 2$. Let $\mathcal{B}$ be any $(n + 1, k)$-code formed from $\mathcal{A}$ by adjoining to $\Omega(\mathcal{A})$ any one of the columns already present in $\Omega(\mathcal{A})$. Let $\mathcal{C}$ be an $(n + 1, k)$-code formed by adjoining to $\Omega(\mathcal{A})$ a column $\mathbf{c}$ not already present in $\Omega(\mathcal{A})$. Then $\mathbf{c}$ can be chosen so that $Q(\mathcal{C}) \geqq Q(\mathcal{B})$ for all $\mathcal{B}$."

This conjecture has been shown not to be true for all $\mathfrak{A}$. E. F. Moore of Bell Telephone Laboratories has constructed a code $\mathfrak{A}$ such that the new code formed by repeating a parity check of $\mathfrak{A}$ is strictly better than any code formed from $\mathfrak{A}$ by adding a new type parity check. The falsity of this conjecture does not preclude the possibility of a thoerem of the sort mentioned in the previous paragraph. One should not expect to pass from a good (n,k)-code to a good $(n + 1, k)$-code in any simple manner: the structure of a best $(n + 1, k)$-code may be quite different from the structure of a best (n,k)-code.

In this connection, we point out that there are many (n,k)-codes that cannot be improved by the addition of a single parity check. This situation obtains whenever the coset leaders of the given code are unique (or, in geometrical terms, when there are no vertices of the n-cube on the boundaries of the maximum-likelihood regions). Adding a single parity check to such a code to form an $(n + 1, k)$-code leaves the value of Q unaltered.

The notions of addition and multiplication for group codes can be easily generalized to hold for block codes. How much of the theory developed remains in this case?

The foregoing are but a few of the many questions that arise naturally from this work. Most of them have not yet been investigated in any detail. We have, it is clear, raised more questions than we have answered. Perhaps this is inherent in the nature of research.

ACKNOWLEDGMENTS

Much of the work reported here was done during the Spring of 1959 while the author was a visiting professor at the University of California in Berkeley. He is indebted to his many friends in the Electrical Engineering Department there for providing a stimulating atmosphere in which to work, and is particularly indebted to Prof. A. J. Thomasian, with whom he discussed many parts of this work.

The author extends his thanks and admiration to Mrs. W. Mammel of Bell Telephone Laboratories, who by ingenious and unusual programs converted the formulae of Section 2.8 into the tables of Section 1.9 (with some aid from an IBM 704).

REFERENCES

1. Slepian, D., A Class of Binary Signalin8 Alphabets, B.S.T.J., **35**, 1956, pp. 203–234.
2. Fontaine, A. B., and Peterson, W. W., Group Code Equivalence and Optimum Codes, I.R.E. Trans., **IT-5**, 1959, pp. 60–70.

3. Riordan, J., The Combinatorial Significance of a Theorem of Pólya, J. Soc. Ind. & Appl. Math., **5**, 1957, p. 225–237.
4. Dickson, L. E., *Linear Groups*, Dover Publications, New York, 1958.
5. Elspas, B., Autonomous Linear Sequential Networks, I.R.E. Trans., **CT-6**, 1959, pp. 45–60.
6. Church, R., Tables of Irreducible Polynomials, Ann. Math., **36**, 1935, pp. 198–209.
7. Littlewood, D. E., *Theory of Group Characters and Matrix Representations of Groups*, Clarendon Press, Oxford, 1950, pp. 166–167.
8. Tucker, A. W., Combinatorial Equivalence of Matrices. mimeographed notes, Princeton Univ., Princeton, N. J.

Capacity of a Burst-Noise Channel

By E. N. GILBERT

(Manuscript received March 15, 1960)

A model of a burst-noise binary channel uses a Markov chain with two states G and B. In state G, transmission is error-free. In state B, the channel has only probability h of transmitting a digit correctly. For suitably small values of the probabilities, p, P of the B → G and G → B transitions, the model simulates burst-noise channels. Probability formulas relate the parameters p, P, h to easily measured statistics and provide run distributions for comparison with experimental measurements. The capacity C of the model channel exceeds the capacity C(sym. bin.) of a memoryless symmetric binary channel with the same error probability. However, the difference is slight for some values of h,p,P; then, time-division encoding schemes may be fairly efficient.

I. INTRODUCTION

In information theory the symmetric binary channel is the classical model of a noisy binary channel. This channel generates a sequence of binary noise digits z_n, which it adds (modulo 2) to input digits x_n to produce output digits $y_n = x_n + z_n$. The symmetric binary channel is memoryless; a sequence of independent trials produces the noise digits z_n. Each trial has the same probability $P(1)$ of producing an error and probability $1 - P(1) = P(0)$ of no error. The capacity $C(\text{sym. bin.})$ of this channel is well known (see Shannon[1]):

$$C(\text{sym. bin.}) = 1 + P(0) \log_2 P(0) + P(1) \log_2 P(1).$$

Channels with memory occur in practice. If radio static or switching transients produce the noise, the errors group into isolated bursts (several errors close together). Independent trials fail to simulate such a burst-noise. Section II of this paper presents a model of a burst-noise channel that is simple enough to permit calculation of the channel capacity C (see Sections III and VI). Sections IV and V give run distributions, the covariance function and other probability formulas as aids to testing the model's applicability and to picking model parameters which match measured statistical data.

Of all binary channels with a given error probability $P(1)$, the symmetric binary channel has least capacity. Indeed, if an encoding for signaling over the symmetric binary channel at a rate R is known, then N sources can use this encoding in time-division multiplex at rates R/N, each over a burst-noise channel. Here, N must be large enough so that noise digits N apart are nearly independent. Time division protects against other noise patterns besides bursts; still less redundant schemes are possible. The possible increase in signaling rate $C - C(\text{sym. bin.})$ will be seen to be often surprisingly small (see Fig. 4).

II. THE MODEL

A Markov chain with two states can be used to generate bursts. The two states will be called G (for good) and B (for bad or for burst). In state G the noise digit is always $z_n = 0$. In state B a coin is tossed to decide whether z_n will be 0 or 1.

The coin-tossing feature is included because actual bursts contain good digits interspersed with the errors. In the formulas that follow a biased coin is allowed (probability h of making no error in state B). All computations given here take $h = 0.50$, which seems a reasonable value.

After producing the noise digit z_n, the Markov chain makes a transition to prepare for z_{n+1}. To simulate burst noise, the states B and G must tend to persist; i.e., the transition probabilities $P = \text{Prob}(G \rightarrow B)$ and $p = \text{Prob}(B \rightarrow G)$ will be small and the probabilities $Q = 1 - P$, $q = 1 - p$ of remaining in G and B will be large. Fig. 1 is a transition diagram for the Markov chain.

Runs of G will alternate with runs of B. The run lengths have geometric distributions with mean $1/P$ for the G-runs and mean $1/p$ for

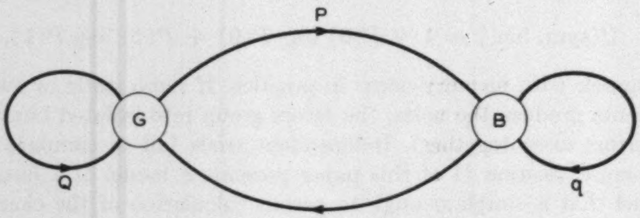


Fig. 1 — Transition diagram for the Markov chain.

the B-runs. The geometric distribution of G-runs seems reasonable. If the various clicks, pops and crashes, which might cause errors on a real channel, are not related to one another, then the times between such events will have the geometric distribution (see Feller,[2] Section XIII.9). Only mathematical simplicity justifies the geometric distribution of B-runs; one might construct more accurate models. Section III mentions one way of elaborating this one; however, complicated models may be useless without adequate statistical data to determine all the model parameters. Section V will illustrate some of the difficulties in determining just the three parameters P, p and h.

The following 500 digits form a typical sample of burst-noise with parameters $P = 0.03$, $p = 0.25$, $h = 0.5$, produced by using random numbers:

$$0^{62}110^{17}10^{46}110101110^{11}110^{15}10^{42}10^{28}110^{90}10^{37}$$

$$110^{5}10010^{35}1011010^{23}110^{4}10^{18}10^{15}11011101101110^{5}.$$

The exponents are run lengths; i.e., 0^{62} denotes a run of 62 consecutive zeros. As expected, long runs of good digits separate the bursts.

The 500-digit sample illustrates the impossibility of reconstructing the sequence of states from the sequence of digits. In portions of some of the long runs of zeros, the Markov chain was in state B; this went unnoticed because the coin tosses produced only zeros. The sample also contains one burst $110^{4}1$ in which a short sojourn into state G produced three of the four zeros.

The fraction of time spent in state B is $P(\mathrm{B}) = P/(p + P)$. Since errors occur only in state B, and then just with probability $1 - h$, the error probability is

$$P(1) = (1 - h)P\,(\mathrm{B}) = (1 - h)\,\frac{P}{p + P}. \qquad (1)$$

III. THE CAPACITY

Let H denote the entropy of the sequence of noise digits $\cdots, z_1, z_2, \cdots$. For all inputs x to the burst-noise channel, the conditional entropy, $H_x(y)$, of the output y knowing the input x is the same:

$$H_x(y) = H.$$

A simple argument then shows that the capacity C of the burst-noise channel is $C = 1 - H$ (a monogram source with probabilities 0.5 for 0 and 0.5 for 1 attains the rate C).

Shannon[1] (Section 7) gives a simple way of computing an entropy H from state probabilities [P(G), P(B) here] and transition probabilities. McMillan[3] (Section 2.0) notes that this result tacitly assumes that the state sequence is reconstructible from the digit sequence. Since a reconstruction is impossible here, H has a more complicated formula.

A definition of H is

$$H = \lim_{N\to\infty} \sum_{z_i=0,1} P(z_1, \cdots, z_N) h(z_1, \cdots, z_N), \tag{2}$$

with

$$h(z_1, \cdots, z_N) = -\sum_{z_{N+1}=0}^{1} P(z_{N+1} \mid z_1, \cdots, z_N) \log_2 P(z_{N+1} \mid z_1, \cdots, z_N). \tag{3}$$

If $z_i = 1$, the corresponding state is certainly B and

$$P(z_{i+1}, \cdots, z_{i+j} \mid z_1, \cdots, z_{i-1}, 1) = P(z_{i+1}, \cdots, z_{i+j} \mid 1) \tag{4}$$

follows for all $j \geqq 1$. Then,

$$P(z_{N+1} \mid z_1, \cdots, z_{i-1}, 1, z_{i+1}, \cdots, z_N) = P(z_{N+1} \mid 1, z_{i+1}, \cdots, z_N)$$

follows and also

$$h(z_1, \cdots, z_{i-1}, 1, z_{i+1}, \cdots, z_N) = h(1, z_{i+1}, \cdots, z_N).$$

Thus, just the number of consecutive zeros at the end of the block $(z_1, \cdots, z_N)$ determine $h(z_1, \cdots, z_N)$ completely. Each of the $2^N h$'s in the sum (2) is one of the $N + 1$ numbers

$$h(1), h(10), \cdots, h(10^k), \cdots, h(10^{N-1}), h(0^N)$$

(again exponents denote run lengths). After using this simplification in (2), summing and letting $N \to \infty$, the result is

$$H = \sum_{K=0}^{\infty} P(10^K) h(10^K). \tag{5}$$

The terms of (5) involve probabilities of runs of zeros. Section IV will give a formula for the conditional probability, $u(K)$, of a run of K or more zeros following a one, that is, $u(K) = P(0^K \mid 1)$. The convention $u(0) = 1$ will be adopted. Then, in (5),

$$P(10^K) = P(1)u(K)$$

[(1) gives $P(1)$]. Also, (3), together with $P(0 \mid 10^K) = u(K + 1)/u(K)$, provides an expression for $h(10^K)$:

$$h(10^K) = -\frac{u(K+1)}{u(K)} \log_2 \frac{u(K+1)}{u(K)} - \left[1 - \frac{u(K+1)}{u(K)}\right] \log_2 \left[1 - \frac{u(K+1)}{u(K)}\right]. \tag{6}$$

Using (6), the terms of (5) rearrange into

$$C = 1 + P(1) \sum_{K=0}^{\infty} v(K) \log_2 v(K), \tag{7}$$

where $v(K) = u(K) - u(K+1)$. Section IV contains formulas for $v(K)$. Although (7) seems simpler than (5) and (6), it converges slowly. In Section V the computation method uses a modification of (5) and (6).

Note that $v(K) = P(0^K1 \mid 1)$. Another derivation of (7) proceeds by showing that the noise sequence consists of successive blocks of digits of the form $1, 01, 001, \cdots, 0^K1, \cdots$, chosen independently, and with probability $v(K)$ for the block 0^K1. Then $-\sum v(K) \log_2 v(K)$ is the information per block and $P(1)$ is the average number of blocks per digit.

Equations (5), (6) and (7) apply to certain other channels. These formulas followed just from (4), which holds whenever the lengths of successive runs of zero are independent. Whenever such independence can be assumed, a more elaborate model might use $v(0), v(1), v(2), \cdots$, directly as parameters. Then $P(1)$ in (7) is

$$P(1) = \left[\sum_{K=0}^{\infty} (K+1) v(K)\right]^{-1}.$$

As a check, the symmetric binary channel has $v(K) = P(1)[P(0)]^K$ and (7) sums to C(sym. bin.).

IV. PROBABILITIES

Recurrent events theory (Feller,[2] Section XIII) provides some probabilities needed in Sections V and VI.

4.1 *Recurrence Times for State B*

Let f_K denote the conditional probability, in state B, that the first return to B will happen at step K:

$$f_K = P(\mathrm{G}^{K-1}\mathrm{B} \mid \mathrm{B}).$$

Then $f_1 = q$, $f_2 = pP$ and $f_K = pQ^{K-2}P$ for $K \geqq 2$. It is convenient to make these probabilities the coefficients of a generating function $F(t)$ of recurrence time probabilities:

$$F(t) = \sum f_K t^K = qt + \frac{pPt^2}{1 - Qt}. \tag{8}$$

For example, the probability $f_K^{(m)}$ that the mth return to B happens at step K has the generating function

$$\sum_{K=1}^{\infty} f_K^{(m)} t^K = [F(t)]^m. \tag{9}$$

The probability of no return to B in k steps is pQ^{k-1}. Then the probability $s(K,m)$ of exactly m returns to B in K steps (but not necessarily a return on step K) is

$$s(K,m) = f_K^{(m)} + \sum_{K=1}^{K-m} f_{K-k}^{(m)} pQ^{k-1}.$$

The corresponding generating function is

$$\sum_{K=1}^{\infty} s(K,m)t^K = \left(1 + \frac{pt}{1 - Qt}\right)[F(t)]^m. \tag{10}$$

4.2. *Recurrence Times for Ones*

Starting from a one (and hence from B), the next one must occur at a return to B, but not necessarily the first return. The probability that the next one occurs at the mth return to B and at step K is

$$h^{m-1}(1 - h)f_K^{(m)}.$$

Then, recurrence time probabilities for ones are

$$v(K - 1) = P(0^{K-1}1 \mid 1) = \sum_{m=1}^{\infty} h^{m-1}(1 - h)f_K^{(m)}.$$

Equation (9) now provides the generating function $V(t) = \sum v(K)t^K$:

$$tV(t) = \frac{(1 - h)F(t)}{1 - hF(t)}. \tag{11}$$

Likewise, the probability $u(K)$ that no one appears in the next K steps is

$$u(K) = \sum_m s(K,m)h^m,$$

which has generating function

$$U(t) = \frac{1 + (p - Q)t}{(1 - Qt)[1 - hF(t)]}. \qquad (12)$$

By (8),

$$U(t) = \frac{1 + (p - Q)t}{D(t)}, \qquad (13)$$

where $D(t) = 1 - (Q + hq)t - h(p - Q)t^2$.

Factor the quadratic $D(t)$:

$$D(t) = (1 - Jt)(1 - Lt),$$

where $2J = Q + hq + \sqrt{(Q + hq)^2 + 4h(p - Q)}$ and L is the same expression with negative square root. Now, (13) becomes

$$U(t) = \frac{1 + (p - Q)t}{J - L}\left(\frac{J}{1 - Jt} - \frac{L}{1 - Lt}\right).$$

The coefficient of t^K in the power series for $U(t)$ is

$$u(K) = \frac{(J + p - Q)J^K - (L + p - Q)L^K}{J - L}. \qquad (14)$$

To find a recurrence formula for $u(K)$, write (13) as $D(t)U(t) = 1 + (p - Q)t$ and equate coefficients of t^K:

$$u(K) = (Q + hq)u(K - 1) + h(p - Q)u(K - 2) \qquad (15)$$

for $K = 2,3,\cdots$. Initial values are

$$u(0) = 1, \qquad u(1) = p + hq.$$

For calculating, (15) is more convenient than (14).

Similar steps lead from (11) to

$$v(K) = \frac{1 - h}{J - L}[(qJ + p - Q)J^K - (qL + p - Q)L^K]. \qquad (16)$$

For $K = 2,3,\cdots$, $v(K)$ also satisfies (15), but with initial values

$$v(0) = (1 - h)q, \qquad v(1) = (1 - \mathrm{h})(pP + hq^2).$$

4.3. *Covariance*

The covariance function of this binary noise is just a joint probability $r(K) = \mathrm{Prob}(z_0 = 1, z_K = 1)$. A formula for the generating function

$R(t) = \sum r(K)t^K$ is

$$\begin{aligned} R(t) &= P(1)\,\{1 + tV(t) + [tV(t)]^2 + \cdots\} \\ &= \frac{P(1)}{1 - tV(t)} \\ &= \frac{P(1)D(t)}{(1-t)[1 + (p - Q)t]}. \end{aligned}$$

The term $P(1)[tV(t)]^m$ in the sum generates the probabilities of finding $z_0 = z_K = 1$, with exactly $m - 1$ of the digits $z_1, \cdots, z_{K-1}$ equal to 1.

An explicit formula for $r(K)$ follows by expanding $R(t)$ in a power series:

$$\begin{aligned} r(0) &= P(1), \\ r(K) &= P(1)^2\left[1 + \frac{p(q - P)^K}{P}\right], \qquad K = 1,2,\cdots. \end{aligned} \tag{17}$$

V. PARAMETER MATCHING

The three parameters p, P, h are not directly observable, so methods of deducing them from statistical measurements must now be considered. We will express p, P, h as functions of three other easily estimated noise parameters. One suitable set of three parameters (involving only trigram statistics) is

$$a = P(1), \qquad b = P(1 \mid 1), \qquad c = \frac{P(111)}{P(101) + P(111)}.$$

Here, c is the conditional probability of finding the place between two ones filled by a one, and it has the expression

$$\frac{(1 - h)q^2}{q^2 + pP}.$$

Solving for p, P, h in terms of a, b, c,

$$\begin{aligned} 1 - p = q &= \frac{ac - b^2}{2ac - b(a + c)}, \\ h &= 1 - \frac{b}{q}, \\ P &= \frac{ap}{1 - h - a}. \end{aligned} \tag{18}$$

If $h = 0.5$ is assumed, then $q = 2b$ and no c measurement is needed.

For illustration, the 500-digit sample in Section II contains thirty-eight 1's, fifteen 11's, seven 101's, and three 111's. Estimates of a, b, c are $a = 38/500$, $b = 15/38$, $c = 3/10$. With these estimates, (18) gives ridiculous parameters (p is negative). The trouble is that 500 digits provide too small a sample. In particular, the estimate $c = 3/10$, based on only 10 observations, is far from the correct value $c = 0.49$. If $h = 0.50$ is assumed, the estimates become $p = 0.21$, $P = 0.036$ (compare with true values $p = 0.25$, $P = 0.03$).

After finding p, P, and h, the results of Section IV suggest comparisons between run measurements and the probabilities $u(K)$ or $r(K)$. Fig. 2 shows curves of some run probabilities $P(10^K) = P(1)u(K)$ (on a log scale) versus K. As shown by (14), these curves straighten out for large K with slopes determined by J.

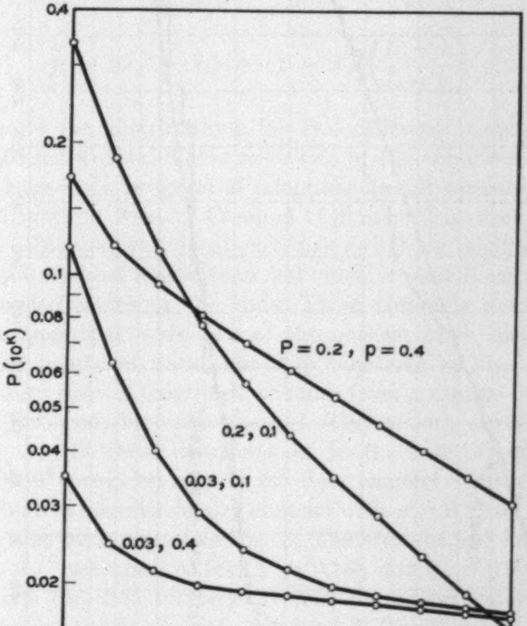


Fig. 2 — Typical run distributions, with $h = \frac{1}{2}$.

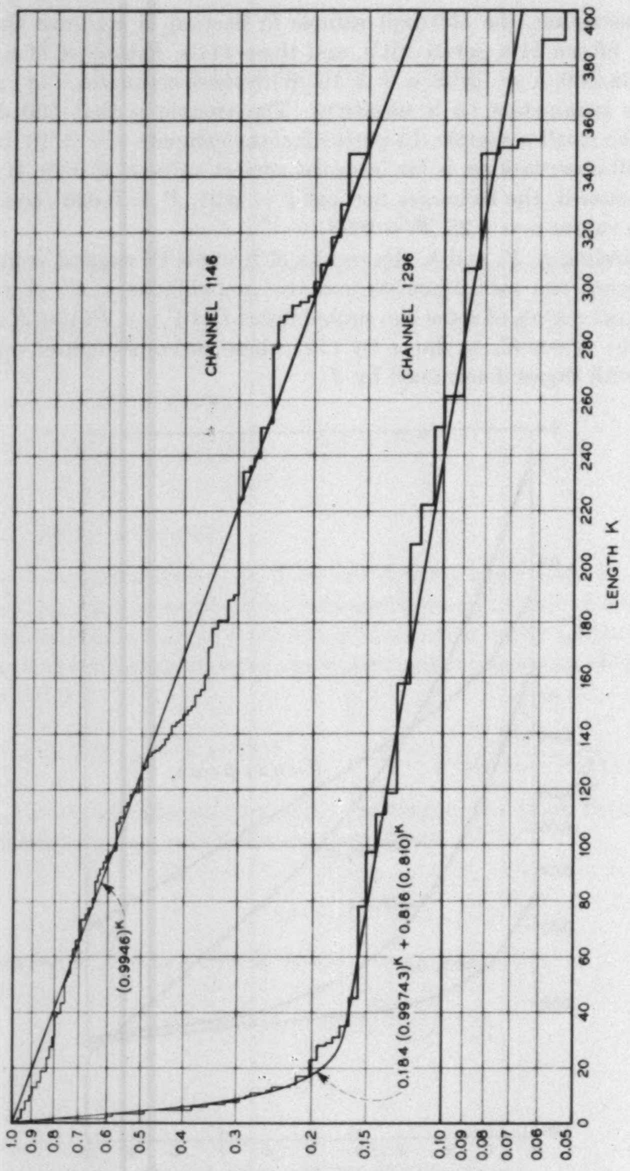


Fig. 3 — Match to experimental run data on telephone channels.

Data on runs of zero can provide another estimate of p, P, h. The fraction of runs of length K or more is an estimate of $u(K)$. By (14), one expects to find constants J, L, A such that

$$u(K) = AJ^K + (1 - A)L^K. \tag{19}$$

These constants are easily found by fitting a curve of the form (19) to the measured run distribution. First, A and J are chosen to give the correct behavior AJ^K for large K. Afterward, L is chosen to improve the fit for small K. Expressions for p, P, h in terms of A, J, L are

$$h = \frac{LJ}{J - A(J - L)},$$

$$P = \frac{(1 - L)(1 - J)}{1 - h},$$

$$p = A(J - L) + (1 - J)\left(\frac{L - h}{1 - h}\right).$$

Fig. 3 shows run distributions for two different telephone circuits transmitting binary data. These were two of the thousands of circuits in a recent large-scale program of telephone circuit measurements (see Alexander, Gryb and Nast.[4] * Channel 1146 carried an exchange call; it used loaded cable and only local exchange switching facilities. Channel 1296 was a toll channel longer than 500 miles; it used K-carrier, a radio path, and loaded cables at the ends. These channels were chosen as examples because they were two of the noisiest cases measured, and thus provided plenty of data. The step functions in Fig. 3 show the fractions of zero runs of lengths K or more from a sample of about 130 consecutive zero runs for each channel. The smooth curves show the curves (19) that fit these distributions. In the case of channel 1146, $u(K) = 0.9946^K$ provided a good fit; then channel 1146 was well approximated by a symmetric binary channel with $p = 0.9946$. The results for channel 1296 look more like Fig. 2. The straight line asymptote is the function AJ^K with parameters $A = 0.184$ and $J = 0.99743$ chosen to approximate the data for large K. The parameter value $L = 0.81$ makes the curve (19) fit the data for small K. These values of A, J, L provide the estimates

$$h = 0.84, \qquad P = 0.003, \qquad p = 0.034.$$

---

* The curves appearing in Ref. 4 show only combined data from hundreds of channels. Since these channels differ greatly among themselves, the curves in Ref. 4 do not have the form (19).

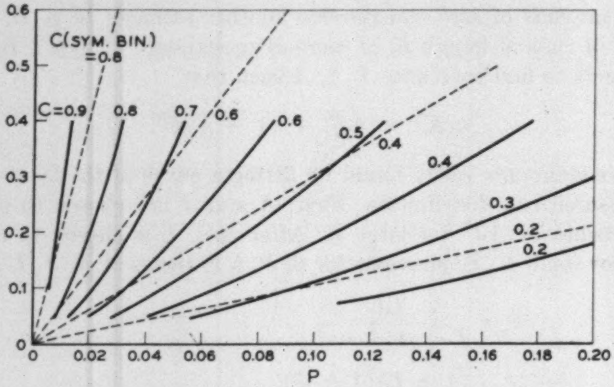


Fig. 4 — Capacities C and C(sym. bin.) as functions of p,P, with $h = \frac{1}{2}$.

The 500-digit sample of Section II provides a run distribution with more statistical fluctuations than in Fig. 3 because of the smaller sample size. The curve fitting yields $A = 0.385$, $J = 0.961$, $L = 0.32$ and $h = 0.432$, $P = 0.047$, $p = 0.232$.

VI. CAPACITY COMPUTATIONS

By (14) and (16), $u(K)$ and $v(K)$ behave like multiples of J^K for large K. In the most interesting cases P is small and J is nearly 1.0 ($J \geqq Q$ always); then (7) converges slowly. However,

$$\frac{u(K+1)}{u(K)} \to J$$

for large K and, by (6),

$$h(10^K) \to -J \log_2 J - (1-J) \log_2 (1-J) = h_0 .$$

Here, $h(10^K)$ approaches its limiting value h_0 rapidly; indeed, $L = Q + hq - J \leqq hq$. When $h = 0.5$, typical values of L are about 0.5 or less, and the L^K term in (14) becomes negligible when K reaches 10 or 15. Thus, the approximation $h(10^K) = h_0$ is good for all $K \geqq K_0$ where K_0 is only moderately large. The corresponding terms of the infinite series (5) sum to

$$\begin{aligned} \sum_{K=K_0}^{\infty} P(10^K)h_0 &= h_0 P(1) \sum_{K=K_0}^{\infty} u(K) \\ &= h_0 \left[1 - P(1) \sum_{K=0}^{K_0-1} u(K) \right]. \end{aligned}$$

The last step used the identity

$$P(1)[u(0) + u(1) + u(2) + \cdots] = 1,$$

which follows from (13) with $t = 1$. Then, the first $K_0 - 1$ terms of (5), together with the correction just derived, suffice to compute C accurately.

Fig. 4 shows contours of constant C and C(sym. bin.) versus p,P for $h = 0.5$. [C(sym. bin.) was computed with $P(1)$ given by (1)]. If the average burst length is not large (p not too small), the difference between the two capacities is slight.

The author is indebted to Miss M. A. Lounsberry for the computations shown in Figs. 2 and 4.

REFERENCES

1. Shannon, C. E., A Mathematical Theory of Communications, B.S.T.J., **27**, 1948; pp. 379; 623.
2. Feller, W., *An Introduction to Probability Theory and Its Applications*, Vol. 1, 2nd Ed., John Wiley & Sons, New York, 1957.
3. McMillan, B., The Basic Theorems of Information Theory, Ann. Math. Stat., **24**, 1953, p. 196.
4. Alexander, A. A., Gryb, R. M. and Nast, D. W., Capabilities of the Telephone Network for Data Transmission, B.S.T.J., **39**, 1960, p. 431.

Automata and Finite Automata

By C. Y. LEE

(Manuscript received March 17, 1960)

Since it is not clear, in general, how an automaton should best be characterized, one of the purposes of this paper is to find ways to go from one characterization to another. In doing so, we have not been completely impartial—the programming approach has been emphasized more than the others. There are perhaps two reasons for this emphasis: First and the more obvious one is the closeness between theoretical programming discussed here and programming of digital computers. Secondly, the programming approach has provided a way of looking at automata that seems to make certain ideas less obscure—the construction of a universal program in Section III of this paper is one such example. In the theory of finite automata, Theorem 3 is an attempt to unify the ideas of complete and partial automata, which have generally been treated separately in the past.

I. INTRODUCTION

The invention of modern computers seems to have been anticipated by many years by Turing.[1] Yet it is remarkable how little the progress of computers has been influenced by Turing's work. There is, perhaps, a basic difference in viewpoint that may account for this lack of convergence. Turing looked at machines from the point of view of their internal behavior. Although Turing originated the concept of universal machines, his idea seems to correspond much closer to that of our special-purpose machines. Every machine, by virtue of its state description, performs a specific task; a machine is altered only if its internal structure is altered. Computers, on the other hand, are generally specified in terms of their external capabilities. Their internal structure remains more or less fixed once they come into being. A computer is then a universal machine in disguise, and every Turing machine corresponds to a particular computer program. One may therefore study the behavior and structure of programs rather than work with states.

The first step in this direction was perhaps taken by Wang,[2] who based his ideas of machines on a computer (which he called a B-machine) that had four kinds of instructions: move to the right or left; mark; transfer conditionally. A B-machine is close to the ultimate in simplicity, but is still capable of computing everything that a Turing machine is capable of and, with a suitable program, is capable of being universal.

As a model, B-machines are attractive because of their intrinsic simplicity. On the other hand, because a B-machine does not have the ability to erase, it is very difficult to write even fairly simple programs without having to work out intricate details. In this paper we have, therefore, introduced a modified B-machine—one which is empowered with the ability to erase. We have called a machine of this kind a W-machine.

The similarities and differences between W-machines and two-symbol Turing machines are shown in Sections II and III. In Section IV we describe the construction of a universal W-machine to show the kinds of techniques involved in W-machine programming. It may be interesting to note here that, once a few useful subprograms are written, the main linkage program takes but a few instructions. Because of its simplicity, one may suspect that it is harder to construct sophisticated combinatorial or symbol-operation kinds of programs on a W-machine than it is on a more complex computer. But we would not be surprised if such a suspicion turns out to be groundless; what makes a W-machine a poor computer may well be only its disregard for time.

The subfamily of W-machines in which each machine has a bounded memory constitutes the family of finite automata. Because finite automata are abstract models of sequential switching circuits, there has been much current interest in their behavior. As a result, there have been a number of approaches to problems in connection with finite automata. In Section V it is shown that finite automata may be characterized by the deletion of one of the five kinds of W-machine instructions. There is thus a program analog of finite automata.

In Section VI the relation between finite automata and sets of input sequences is discussed. Among other things we present within our framework a result of Kleene[3] that makes it possible to represent finite automata by algebraic-like expressions. This characterization seems very natural in many ways, except that the expressions can easily get very lengthy. The problem of how best to handle these expressions appears very intriguing and, as far as we know, is quite open.

II. TURING MACHINES

A machine will be called an A-machine if it consists, aside from its control mechanism, of the following:

i. A one-way potentially infinite tape (say infinite to the right) divided into squares. Each square can either be marked (having in it the symbol 1) or erased (having in it the symbol 0), and

ii. A reading and writing head that scans some square of the tape at any discrete moment of time. Since the tape is finite to the left, the machine is assumed to stop if the read-write head is ordered to go to the left of the leftmost square of the tape.

The content c_0 of the tape of the A-machine previous to the initial moment of time, consisting of a finite sequence of zeros and ones, is called the (tape) *input* to the A-machine. As time advances, the tape content would change unless some stable condition is reached, so that we would get a sequence c of tape contents $(c_0, c_1, \cdots)$, where c_j is a later tape content than c_i if $i < j$, and where $c_i \neq c_{i+1}$. The sequence c is called the *external behavior* of the A-machine relative to the tape input c_0. Two A-machines are said to be *completely equivalent* if they have identical external behaviors relative to all tape inputs. That is, two A-machines are completely equivalent if they cannot be distinguished by anyone observing just the sequence of tape contents.

The idea of complete equivalence is too stringent at times. If an A-machine is used to compute values of a function, what the machine does while it is processing its data is, in a sense, irrelevant as long as the final answer turns out to be the desired answer. We will, later on, also consider a less stringent type of equivalence.

The fact that an A-machine has a potentially infinite tape implies that it has an indefinitely large memory. It might be helpful to keep the notion that the tape is finite at any moment, but that at any moment a finite amount of blank tape may be added to the right whenever such a demand arises. In the same way, it is helpful to note that every input is a finite sequence of zeros and ones. We will, however, speak of the *null input*, meaning a string of zeros indefinitely long. The null input corresponds to an indefinitely long blank tape.

We will consider the following model of a Turing machine, hereafter called a T-machine, as one of the A-machines. In addition to being an A-machine, it has k active internal states $q_1, q_2, \cdots, q_k$ and an inactive state q_0 in which the machine is assumed to stop. The machine can have one of the following combination of actions: erase or mark the square under scan; move the read-write head one square to the left or one square to the right; go into some state q_j. A T-machine is completely specified if its combination of actions is specified for every state of the machine and each of the two symbols under scan, and if the initial state and the initial square under scan are given.

For instance, the following one-state (i.e. one active state) T-machine, if started initially scanning a square in the interior of its tape, will have its read-write head swinging back and forth, changing ones to zeros while going in one direction and changing zeros to ones while going in the other direction. The read-write head will either proceed indefinitely to the right or will eventually stop at the leftmost square. In this and later description of A-machines, we will use the letter m to denote the action of marking the square under scan; e for the action of erasing the square under scan; $+$ for the action of moving the read-write head one square to the right of the square under scan; and $-$ for the action of moving the read-write head one square to the left:

State	Symbol	
	0	1
$^*q$	$m, +, q$	$e, -, q$

Here q designates the single active state of the T-machine, and * denotes the fact that q is also the initial state of this machine. If the square under scan is not marked, a mark is put in it, the read-write head moves one square to the right, and the machine returns to state q. If the square under scan is marked, it is then erased, the read-write head moves one square to the left, and the machine again returns to state q.

From now on, we will at times use the notation q_i ; m or e, $+$ or $-$, q_j ; m or e, $+$ or $-$, q_k for each combination of actions of any T-machine. Thus, the combination of actions of the one-state T-machine in question can be written: $q; m, +, q; e, -, q$.

III. W-MACHINES

A W-machine is an A-machine together with a program made up of an ordered list of the following five types of base instructions: (a) e: erase the square under scan; (b) m: mark the square under scan; (c) $+$: move the read-write head one square to the right; (d) $-$: move the read-write head one square to the left; and (e) $t(A)$: transfer to program address A if the square under scan is marked, otherwise transfer to the next program address on the ordered list. These base instructions are executed in order by a control mechanism. The initial program address and the initial square under scan are given.

A program of a W-machine consisting of all base instructions with each instruction having a separate address is called a *base program.* Let us consider a W-machine completely equivalent to the one-state T-machine illustrated earlier. The base program for this machine is

1. *t*7
2. *m*
3. +
4. *t*7
5. *m*
6. *t*2
7. *e*
8. −
9. *t*7
10. *m*
11. *t*2.

We note that the instructions in the program refer to only two addresses, address 2 and address 7. The program may therefore be equally well written

1. *t*3
2. *m*, +, *t*3, *m*, *t*2
3. *e*, −, *t*3, *m*, *t*2,

where the instructions contained in one line are understood to be executed consecutively. This notation simplifies the writing of W-machine programs and will be used in this paper wherever it is convenient to do so.

A base program of a W-machine is said to be *minimal* if there is no W-machine completely equivalent to it with fewer base instructions in its program. In order not to have to consider special cases later, let us agree at this stage to rule out certain trivial redundancies in W-machine programs. Consider two W-machines, W_1 and W_2, as follows:

Machine W_1

1. *m*
2. +
3. *t*1

Machine W_2

1. *m*
2. *e*
3. *m*
4. +
5. *t*1.

Machines W_1 and W_2 are not completely equivalent, since they have nonidentical external behavior. The difference is, however, of a minor nature. We will therefore agree that, whenever a W-machine program contains consecutive instructions

A. *e* or *m*
$A + 1$. *e* or *m*
⋮
$A + i$. *e* or *m*,

only the last instruction (in address $A + i$) will be retained, and the others will be deleted. Furthermore, if it should become necessary to mark and erase a square in succession, the final symbol in that square will be accepted as the output symbol for that moment.

The fact that a base program is minimal itself implies that the base program cannot contain certain subprograms.

Lemma 1:

Let P be a minimal base program of a W-machine. Then P cannot have two consecutive addresses A and $A + 1$ having in them the following base instructions:

(i) A. $t(B)$
$A + 1$. $t(C)$;

(ii) A. e
$A + 1$. $t(C)$;

(iii) A. e
$A + 1$. e;

(iv) A. m
$A + 1$. m.

Proof: In (i) and (ii), if address $A + 1$ is never referred to, P cannot be minimal since the $(A + 1)$th instruction can be deleted. On the other hand, if there is some instruction $t(A + 1)$ in P, such an instruction can be changed to $t(C)$, again making the $(A + 1)$th instruction superfluous. This proves (i) and (ii); (iii) and (iv) are obvious, and the lemma follows.

Theorem 1:

I. Given a W-machine having b base instructions, there is a completely equivalent T-machine with not more than b states.

II. Given a T-machine with s states, there is a completely equivalent W-machine with not more than $10s + 1$ base instructions.

Proof: Let a W-machine with b base instructions be given. That there is a completely equivalent T-machine is clear. It remains for us to show for part I of the theorem that b states would suffice.

Let P be a minimal base program for the W-machine and A be the initial address of P. Then, by Lemma 1, the base instructions in addresses A and $A + 1$ are one of the following:

(i) A. $t(B)$
$A + 1$. e;

(ii) A. $t(B)$
$A + 1$. m;

(iii) A. $t(B)$
$A + 1$. $+$ or $-$;

(iv) A. m or e
$A + 1$. $+$ or $-$;

(v) A. $+$ or $-$
$A + 1$. m or e;

(vi) A. $+$ or $-$
$A + 1$. $+$ or $-$.

In (i), (ii) and (iii), we assert that the base instruction in address B can be made one of the following:

$$(a)\ B.\quad e \qquad \text{or} \qquad (b)\ B.\quad + \text{ or } -.$$

This is true because, if the instruction in B should be $t(C)$ and the instruction in C should be $t(D)$ and so on, then at some point in the chain, say address E, the instruction must be a nontransfer instruction, for otherwise the program would not have been minimal. We may then replace the instruction $t(B)$ in A by $t(E)$. On the other hand, if the instruction in B should be m, then the instruction in A could have been replaced by:

$$A.\quad t(B + 1);$$

and the assertion follows.

In (i) and case (a), by Lemma 1, the base instructions in addresses $A + 2$ and $B + 1$ must be $+$ or $-$. Thus, address A can be associated with a T-machine state

$$q(A);\quad e, + \text{ or } -, q(A + 3);\quad e, + \text{ or } -, q(B + 2).$$

Similarly, in case (b) address A can be associated with a T-machine state

$$q(A);\quad e, + \text{ or } -, q(A + 3);\quad m, + \text{ or } -, q(B + 1).$$

It should be noted that a T-machine state may replace more than just address A. For example, in (i) case (a) the T-machine state replaces the five addresses A, $A + 1$, $A + 2$, B and $B + 1$ if none of these addresses is referred to elsewhere in the program. Therefore, in going from a W-machine to a T-machine as described by the procedure outlined here, the T-machine will in general have fewer than b states.

In (ii), the $(A + 2)$th instruction can be either

$$A + 2.\quad + \text{ or } - \qquad \text{or} \qquad A + 2.\quad t(C).$$

The former is no different from (i). In the latter, the instruction in address C can be made one of the following:

$$C.\quad e \qquad \text{or} \qquad C.\quad + \text{ or } -.$$

The T-machine states to be associated with address A in case (a) corresponding to these two subcases are respectively

$$q(A);\quad e, + \text{ or } -, q(C + 2);\quad e, + \text{ or } -, q(B + 2),$$

and

$$q(A);\quad m, + \text{ or } -, q(C + 1);\quad e, + \text{ or } -, q(B + 2);$$

and in case (b) are respectively

$$q(A); \quad e, + \text{ or } -, q(C + 2); \quad m, + \text{ or } -, q(B + 1),$$

and

$$q(A); \quad m, + \text{ or } -, q(C + 1); \quad m, + \text{ or } -, q(B + 1).$$

Case (iii) is similar to (i). In (iv) and (v), the two addresses, A and $A + 1$, can obviously be associated with a single T-machine state. In (vi), each address A or $A + 1$ may be associated with a single T-machine state. Therefore, there is a completely equivalent T-machine with not more than b states and part I of the theorem follows.

To prove part II, let a T-machine with s states be given with states q_i, $i = 1, 2, \cdots, s$:

$$q_i; \quad a_i(0), b_i(0), q_i(0); \quad a_i(1), b_i(1), q_i(1),$$

where a_i is either m or e and b_i is either $+$ or $-$. Associate with each state q_i two addresses A_i and A_i' of a W-machine:

$$A_i . \quad a_i(0), b_i(0), t[A_i'(0)], m, t[A_i(0)];$$

$$A_i' . \quad a_i(1), b_i(1), t[A_i'(1)], m, t[A_i(1)].$$

Next, if q_j is the initial state of the T-machine, we will add an initial address $A_j - 1$ where we have

$$A_j - 1. \qquad t(A_j').$$

The W-machine so defined is completely equivalent to the T-machine, having exactly $10s + 1$ base instructions. This proves part II of the theorem.

The bound $10s + 1$ on the number of base instructions cannot be lowered if the first address is to be always the initial address of a W-machine program. If we are allowed to begin a program at some intermediate address, the bound $10s + 1$ can be lowered to perhaps $8s + 1$.

From this result, it follows that whatever is true about T-machines is functionally true about W-machines, and conversely. The choice of whether to use the T-machine or the W-machine model is therefore somewhat arbitrary. We have found that the T-machine model is convenient for state description of finite automata (Section V) and the W-machine model more satisfactory for problems involving operations with symbols. The latter contention is illustrated by a universal W-machine described below.

IV. A UNIVERSAL W-MACHINE.

A very interesting result of Turing[1] was his construction of a machine which is capable of imitating any target machine when given a description of the target machine. Such a machine is known as a universal Turing machine. To our knowledge there have been two independent and very ingenious constructions of universal machines which greatly simplified Turing's work. The earlier one is due to Moore[4] and the other to Ikeno.[5] Moore's machine has two symbols and three tapes, and can be reduced to nine states. Ikeno's machine requires six symbols and ten states, giving a state-symbol product of 60.

A word should perhaps be said in regard to the two extra tapes in the case of Moore's machine and the four extra symbols used in Ikeno's machine. In either case, the universal machine is just slightly different from the target machines it imitates. It would be more "authentic" for a universal machine to be immediately within the class of all target machines it imitates. It also seems that it is as direct to construct such an "authentic" machine as otherwise. For these reasons, we include here the construction of a universal W-machine as an example.

Let U denote the universal W-machine to be constructed. Let the squares on the tape of U be divided into two classes: *a*-squares and *b*-squares. If the squares are numbered beginning with one for the left-most square and proceeding to the right, then the *a*-squares are the odd-numbered squares and the *b*-squares the even-numbered squares. The *b*-squares are there to serve as markers. The description of the target machine together with the data occupies only the *a*-squares. In order to clarify coding, the contents of *a*-squares are underlined. Thus $\underline{1}\ \underline{0}$ would mean a mark and a blank in adjacent *a*-squares, the content of the inbetween *b*-square has been left unspecified.

The W-machine instructions are coded as follows:

Instruction	*Code*
* (Stop)	1
+	$\underline{1}\ \underline{1}$
−	$\underline{1}\ \underline{1}\ \underline{1}$
m	$\underline{1}\ \underline{1}\ \underline{1}\ \underline{1}$
e	$\underline{1}\ \underline{1}\ \underline{1}\ \underline{1}\ \underline{1}$
$t(n)$	$\underline{1}\ \underline{1}\ \underline{1}\ \underline{1}\ \underline{1}\ \underbrace{\underline{1}\ \underline{1}\ \cdots\ \underline{1}}_{n \text{ ones}}.$

The program and data of the target W-machine that the universal W-machine U is to imitate occupy only the *a*-squares on the tape of U. The instructions are coded in sequence, with a single blank *a*-square separating adjacent instructions. The data go directly into *a*-squares without modification. There is a single blank *a*-square between the last instruction and the data.

The first two *a*-squares are blank and all *a*-squares to the right of the data are blank. The *b*-squares are all marked except for (a) the first *b*-square, (b) the *b*-square immediately to the right of the data square under scan and (c) all *b*-squares to the right of square x, where x is the *a*-square to the right of the last data *a*-square.

The coding scheme will be made clear by an example. Suppose the program of the target W-machine is

1. $t3$
2. $+$
3. m
4. $t2$,

where the initial address is address 1 and the data are

$$1\ 0\ \overset{\downarrow}{1}\ 1$$

where the third symbol is the initial symbol under scan. In the coded form, the tape of U would have contents

$$\underline{0}\ 0\ \underline{0}\ 1\ \underbrace{\overset{\overset{\text{Begin}}{\downarrow}}{\underline{1}}\ 1\ \underline{1}\ 1\ \underline{1}\ 1\ \underline{1}\ 1\ \underline{1}\ 1\ \underline{1}\ 1\ \underline{1}\ 1\ \underline{1}}_{t3}\ 1\ \underline{0}\ 1\ \underbrace{\underline{1}\ 1\ \underline{1}}_{+}\ 1$$

$$\underline{0}\ 1\ \underbrace{\underline{1}\ 1\ \underline{1}\ 1\ \underline{1}\ 1\ \underline{1}}_{m}\ 1\ \underline{0}\ 1\ \underbrace{\underline{1}\ 1\ \underline{1}\ 1\ \underline{1}\ 1\ \underline{1}\ 1\ \underline{1}\ 1\ \underline{1}\ 1\ \underline{1}}_{t2}\ 1$$

$$\underline{0}\ 1\ \underbrace{\underline{1}\ 1\ \underline{0}\ 1\ \overset{\overset{\text{Symbol under scan}}{\downarrow}}{\underline{1}}\ 0\ \underline{1}}_{\text{Data}}\ 1\ \underline{0}\ 0\ \underline{0}\ \cdots$$

The program for the universal W-machine U is divided into a main program P and a number of subprograms. The various subprograms are designated by symbolic addresses as follows:

RT	One square to the right.
LT	One square to the left.
MK	Mark square under scan.
ER	Erase square under scan.
TR	Transfer if data square under scan is marked. If transfer is effective, go to the beginning of tape and hunt to the right until the correct instruction has been found. Otherwise, go to the next instruction.
RTZ	Right to zero.
LTZ	Left to zero.
RDZ	Right to double zero.
LDZ	Left to double zero.

The program for U begins with the main program P. It first examines the instruction to be carried out. If the instruction should be +, −, *m* or *e*, the program enters subprograms RT, LT, MK or ER respectively. If the instruction should be *t*(*n*), the program enters subprogram TR.

Let us begin with the basic subroutines RTZ, LTZ, RDZ and LDZ:

```
RTZ    1. +2, t1.
LTZ    1. −2, t1.
RDZ    1. +2, t1, +2, t1.
LDZ    1. −2, t1, −2, t1.
```

Next the subprograms TR, RT, LT, MK and ER:

```
TR     1. +, e, RTZ, −, t2, −, t(LT3),
       2. LDZ, +4,
       3. e, +, RTZ, m, +, t4, +, LTZ, −, RDZ, m, t(P),
       4. +, e, LTZ, −, RDZ, m, RTZ, +2, e, LDZ, m, t3.
LT     1. +, e, RTZ, m, −2,
       2. e,
       3. LTZ, m, +, t(P).
RT     1. +, e, RTZ, m, +2, m, t(LT2).
MK     1. +, e, RTZ, −, m, t(ER2).
ER     1. +, e, RTZ, −, e,
       2. −, t(LT3).
```

Finally, the main program P:

```
P      1. +2, t2, *,
       2. +2, t3, RT,
       3. +2, t4, LT,
       4. +2, t5, MK,
       5. +2, t(TR), ER.
```

After the following sequence of instructions of the target machine has been executed:

$$t3,\ m,\ t2,\ +,$$

the tape contents read:

$$\underline{0}\;0\;\underline{0}\;1\;\underbrace{\overset{\substack{\text{Begin}\\ \downarrow}}{\underline{1}}\;1\;\underline{1}\;1\;\underline{1}\;1\;\underline{1}\;1\;\underline{1}\;1\;\underline{1}\;1\;\underline{1}\;1\;\underline{1}}_{t3}\;1\;\underline{0}\;1\;\underbrace{\underline{1}\;1\;\underline{1}}_{+}\;1$$

$$\underline{0}\;1\;\underbrace{\overset{\substack{\text{Begin}\\ \text{instruction } m\\ \downarrow}}{\underline{1}}\;1\;\underline{1}\;1\;\underline{1}\;1\;\underline{1}}_{m}\;1\;\underline{0}\;1\;\underbrace{\underline{1}\;1\;\underline{1}\;1\;\underline{1}\;1\;\underline{1}\;1\;\underline{1}\;1\;\underline{1}\;1\;\underline{1}}_{t2}\;1$$

$$\underline{0}\;1\;\underbrace{\underline{1}\;1\;\underline{0}\;1\;\underline{1}\;1\;\overset{\substack{\text{Symbol}\\ \text{under scan}\\ \downarrow}}{\underline{1}}}_{\text{Data}}\;0\;\underline{0}\;0\;\underline{0}\;\cdots$$

We will call a target W-machine *admissible* if its read-write head never goes to the left of the leftmost square on tape. Machine U then imitates all admissible target machines and is itself admissible.

It may be interesting to note that the coding for machine U does not make an intrinsic distinction between program and data. The burden of distinguishing which is program and which is data is therefore on the coder.

Using the conversion procedure discussed in the proof of Theorem 1, there is a T-machine completely equivalent to the W-machine U with about 74 internal states.† The program for U itself requires some 125 base instructions. As things go, it is not impossible for someone to improve our result to a 50 base instruction universal W-machine or a 25-state universal T-machine or perhaps even better. The answer to the problem of finding a universal machine with the smallest state-symbol product posed by Shannon[6] seems to be quite remote, even for two-symbol machines.

---

† Some of the ideas that resulted in this construction were due to D. Younger, who indicated a possible reduction to a machine of about 56 states.

V. FINITE AUTOMATA

There is a subfamily of T-machines that are abstract models of a class of switching circuits called sequential circuits. The dominant trait of these machines is a strictly limited memory, so that they are called *finite automata*. (These machines are also known as sequential machines.) Because of their limited memory, rather simple tasks lie beyond the reach of finite automata. For instance, there is no finite automaton that, having the null input and ejecting symbols one at a time, will give us the successive digits of π or, for that matter, any number that is not rational. On the other hand, many decision problems become finite problems for finite automata; in fact, in some cases efficient algorithms have been found.

A two-symbol finite automaton consists of

i. A finite number of internal states $q_0, q_1, \cdots, q_n$.

ii. An alphabet of two symbols: $s_0 = 0$, $s_1 = 1$.

iii. A map M whose domain and range are both subsets of the set of state-symbol pairs. If M is defined for a state-symbol pair (q_i, s_j), then $M(q_i, s_j)$ is another pair (q_k, s_r). The symbol s_j is called an *input symbol*. The symbol s_r is called an *output symbol*, and is completely determined by q_i; that is, s_r is independent of the input symbol s_j.

iv. An initial state q_0, which can reach every state q_i, $0 < i \leqq n$, via some suitable input sequence of symbols.

In the definition of a finite automaton given above, we included those automata in which the map M may be undefined for some state-symbol pairs (q_i, s_j). We will call such automata *partial automata*. Partial automata in the past have been treated somewhat differently from complete automata. By considering certain input sequences called acceptable sequences, we will be able to treat partial and complete automata on a uniform basis.

5.1 *Finite Automata and W*-Machines.*

In the beginning of this section we mentioned that finite automata can be regarded as a subfamily of T-machines, and hence as a subfamily of W-machines. Let us call a W-machine a W*-machine if the base program of the W-machine does not contain the instruction "$-$"; that is, if the read-write head of the W-machine never moves to the left. We will see that, by suitable interpretation of inputs and outputs, every finite automaton is completely equivalent to some W*-machine and, furthermore, that every W*-machine differs from some finite automaton by at most a unit of delay in the output.

Let S be a finite automaton; S may then be considered as a T-machine in the following sense: An input sequence of symbols to S corresponds to having this sequence of symbols on the tape of the T-machine, beginning with the initial input symbol on the leftmost square of the tape. In operation, the T-machine begins by scanning the initial square, writes the output symbol on the square being scanned, moves one square to the right and goes into its next state. At any moment, therefore, the previous output is contained in the square just to the left of the read-write head, and the present input is contained in the square directly under the read-write head. In this way the read-write head of the T-machine never moves to the left. It follows from Theorem 1, therefore, that there is a W-machine whose program consists of no base instruction of the form "$-$" and is such that this W-machine and the T-machine are completely equivalent.

Conversely, suppose a W*-machine is given. By Theorem 1, there is a T-machine completely equivalent to this W*-machine such that its read-write head never travels to the left. Such a T-machine may not be in the form of a finite automaton since its output symbol may be a function of both the input symbol and the current state of the machine. We wish to show therefore that such a T-machine differs from a finite automaton by at most a unit of delay in the output.

Consider a T-machine whose read-write head never travels to the left. It then consists of states of the following kind:

$$q_i; \quad a_i, +, q_j; \quad b_i, +, q_k,$$

where a_i and b_i are either e or m. In the particular case $a_i = b_i$ for some i, the output becomes in no way dependent upon the input. We will therefore consider only those states q_i for which $a_i \neq b_i$.

Let us now form a new T-machine by splitting each such state q_i of the original T-machine into two states, q_{i0} and q_{i1}, such that we have for the new machine,

$$q_{i0}; \quad a_i, +, q_{j0}; \quad a_i, +, q_{j1}$$

and

$$q_{i1}; \quad b_i, +, q_{k0}; \quad b_i, +, q_{k1},$$

and, if q_0 should be the initial state of the original T-machine, add a new state q_0^* as the initial state of the new machine:

$$q_0^*; \quad e, +, q_{00}; \quad e, +, q_{01}.$$

In operation, the new machine imitates the original machine faithfully, except that the output of the new machine is delayed by a unit of time; that is, the present output of the new machine is the previous output of the original machine. We have therefore

Theorem 2: Every finite automaton with s states is completely equivalent to a W*-machine with not more than $10s + 1$ base instructions. Every W*-machine with b base instructions differs from a finite automaton of not more than $2b + 1$ states by at most one unit of delay in the output.

An Example. Consider the following W*-machine:

1. $t6$
2. $+$
3. m
4. $t8$
5. e
6. $+$
7. $t2$
8. $+$
9. $+$
10. $t5$.

This W*-machine is completely equivalent to a five-state T-machine with initial state q_1:

State	Symbol	
	0	1
*q_1	$e, +, q_3$	$m, +, q_7$
q_3	$m, +, q_9$	$m, +, q_9$
q_7	$e, +, q_9$	$m, +, q_3$
q_9	$e, +, q_{10}$	$m, +, q_{10}$
q_{10}	$e, +,$ stop	$e, +, q_7$

The T-machine is not in the form of a finite automaton, since its output symbols depend on both the state and the input symbol. Let us therefore split each state whose output symbol is different for different input symbols into two states and, in addition, define a new initial state $q_0{}^*$. The machine then becomes:

State	Symbol	
	0	1
*$q_0{}^*$	$e, +, q_{1.0}$	$e, +, q_{1.1}$
$q_{1.0}$	$e, +, q_3$	$e, +, q_3$
$q_{1.1}$	$m, +, q_{7.0}$	$m, +, q_{7.1}$
q_3	$m, +, q_{9.0}$	$m, +, q_{9.1}$
$q_{7.0}$	$e, +, q_{9.0}$	$e, +, q_{9.1}$
$q_{7.1}$	$m, +, q_3$	$m, +, q_3$
$q_{9.0}$	$e, +, q_{10}$	$e, +, q_{10}$
$q_{9.1}$	$m, +, q_{10}$	$m, +, q_{10}$
q_{10}	$e, +,$ stop	$e, +, q_{7.1}$

This machine is identical with the original W*-machine except that its output symbols are delayed by a unit of time and its initial output symbol is always a zero. For the same input, the sequence of the tape contents of the two machines are therefore not exactly the same; the tape content of the new machine to the left of the read-write head is the tape content of the original machine to the left of the read-write head translated one square to the right. The tape contents of the two machines to the right of the read-write head are, of course, the same.

Since the output of a finite automaton depends only on its state, and since the symbol + is redundant, the state-symbol table of a finite automaton can be simplified. For instance, the nine-state machine given in the example can be given by:

State	Symbol		Output
	0	1	
$*q_0*$	$q_{1,0}$	$q_{1,1}$	0
$q_{1,0}$	q_3	q_3	0
$q_{1,1}$	$q_{7,0}$	$q_{7,1}$	1
q_3	$q_{9,0}$	$q_{9,1}$	1
$q_{7,0}$	$q_{9,0}$	$q_{9,1}$	0
$q_{7,1}$	q_3	q_3	1
$q_{9,0}$	q_{10}	q_{10}	0
$q_{9,1}$	q_{10}	q_{10}	1
q_{10}	stop	$q_{7,1}$	0

For complete automata, except for including the initial state in our model, this description is the same as that given by Moore.[7] In the same way, the description of the five-state T-machine in the example which is completely equivalent to the original W*-machine can also be simplified. We may write

State	Symbol		Outputs	
	0	1		
$*q_1$	q_3	q_7	0	1
q_3	q_9	q_9	1	1
q_7	q_9	q_3	0	1
q_9	q_{10}	q_{10}	0	1
q_{10}	stop	q_7	0	0

where to each state may be associated two output symbols, one for each input symbol. This description is essentially the model of sequential machines used by Huffman[8] and Mealy.[9] It is quite clear from the foregoing that there is a close relationship between these two models, and that one may go freely from one to the other.†

---

† Another way of relating models of finite automata is discussed by Cadden.[10]

5.2 *Finite Automata with a Minimum Number of States*

A problem of interest to switching circuit designers is finding finite automata having a smallest number of states. In relay circuit design, for example, the number of relays needed is usually a monotone function of the number of states the circuit has. For such circuits, therefore, the number of states becomes in a way a measure of cost.

Let A be a partial finite automaton. A finite sequence is said to be an *acceptable sequence* for A if there is an output sequence and a terminating state when this sequence is presented as the input sequence to A, with A beginning in its initial state. We will call the set of all acceptable sequences for A the *acceptable set* for A and denote this set by $R(A)$. Now let A and B be two partial finite automata and let the intersection $R_{AB} = R(A) \wedge R(B)$ be called the *common acceptable set* for A and B. Then A and B are said to be *completely equivalent* with respect to R_{AB} if, for all input sequences belonging to R_{AB}, A and B give identical output sequences. If R is a subset of R_{AB}, then equivalence of A and B with respect to R is defined similarly. It is clear that this definition of complete equivalence is the same as that given before for T- and W-machines, except the input sequences are now restricted to just the acceptable sequences.

As an example consider A and B defined as

A:

State	Symbol		Output
	0	1	
$^*a_0$	a_1		1
a_1		a_1	0

B:

State	Symbol		Output
	0	1	
$^*b_0$		b_1	1
b_1		b_2	0
b_2	b_1		0

The acceptable set $R(A)$ for A is the set of all finite sequences $\{0, 01, 011, 0111, \cdots\}$ and the acceptable set $R(B)$ for B is the set of finite sequences $\{1, 11, 110, 1101, 11010, 110101, \cdots\}$. There is no sequence that is acceptable to both A and B. The common acceptable set R_{AB} is therefore empty.

Theorem 3: Let A and B be two partial finite automata with a and b states respectively, where $a, b > 1$. Let R_{AB} be the common acceptable set for A and B and let $R_{AB}(l)$ be the subset of R_{AB} such that every sequence in $R_{AB}(l)$ is of length $\leq l$. Then A and B are completely equivalent with respect to R_{AB} if and only if they are completely equivalent with respect to $R_{AB}(l)$ for $l = ab - 2$.

Before going through the proof, it would be helpful to discuss some notations that will be used later. If A is a finite automaton, its initial state will be denoted by a_0. If q is any state of A, we will denote the output symbol associated with the state q by $\omega(q)$. Moreover, let $s = (s_0, s_1, \cdots, s_{m-1})$ be an acceptable input sequence for A. Then we will at times let a_i stand for the state reached by A after receiving the ith symbol of s. It will be convenient here to speak of the motion diagram for A:

$$\begin{tikzcd}
\text{Input symbols:} & & s_0 \arrow[dr] & & s_1 \arrow[dr] & & \cdots & & s_{m-1} \arrow[dr] & \\
\text{Transition of states of } A\text{:} & a_0 \arrow[ur, dash] & & a_1 \arrow[ur, dash] & & a_2 & \cdots & a_{m-1} \arrow[ur, dash] & & a_m .
\end{tikzcd}$$

Proof: The theorem is clear in one direction. In the other direction, let A and B be completely equivalent with respect to the set $R_{AB}(ab - 2)$; that is, A and B will give identical output sequences to every commonly acceptable sequence of length not greater than $ab - 2$.

Let us now suppose that there is a common acceptable sequence $s = (s_0, s_1, \cdots, s_{m-1})$ of minimum length m where $m > ab - 2$ such that in the motion diagram for A and B we have

$$\begin{tikzcd}
& s_0 \arrow[dr] & & s_1 \arrow[dr] & & \cdots & & s_{m-1} \arrow[dr] & \\
a_0 \arrow[ur, dash] & & a_1 \arrow[ur, dash] & & a_2 & \cdots & a_{m-1} \arrow[ur, dash] & & a_m \\
& {} \arrow[dr] & & {} \arrow[dr] & & & & {} \arrow[dr] & \\
b_0 \arrow[ur, dash] & & b_1 \arrow[ur, dash] & & b_2 & \cdots & b_{m-1} \arrow[ur, dash] & & b_m
\end{tikzcd}$$

where

$$\omega(a_i) = \omega(b_i) \quad \text{for} \quad i = 0, 1, \cdots, m - 1 \quad \text{but} \quad \omega(a_m) \neq \omega(b_m).$$

There are now two cases to consider. The case $m > ab - 1$ is simpler and will be left to the reader.

Let us therefore assume $m = ab - 1$. In the motion diagram above, we have then exactly ab pairs of states: $(a_0, b_0), (a_1, b_1), \cdots, (a_m, b_m)$. First, suppose that these ab pairs are not distinct; that is, suppose $(a_i, b_i) = (a_j, b_j)$ for some $0 \leqq i < j \leqq m$. The motion diagram then becomes

$$\begin{tikzcd}
& s_0 \arrow[dr] & & \cdots & & s_i \arrow[dr] & & \cdots & & s_j \arrow[dr] & & \cdots & & s_{m-1} \arrow[dr] & \\
a_0 \arrow[ur, dash] & & a_1 & \cdots & a_i \arrow[ur, dash] & & a_{i+1} & \cdots & a_j \arrow[ur, dash] & & a_{j+1} & \cdots & a_{m-1} \arrow[ur, dash] & & a_m \\
& {} \arrow[dr] & & & & {} \arrow[dr] & & & & {} \arrow[dr] & & & & {} \arrow[dr] & \\
b_0 \arrow[ur, dash] & & b_1 & \cdots & b_i \arrow[ur, dash] & & b_{i+1} & \cdots & b_j \arrow[ur, dash] & & b_{j+1} & \cdots & b_{m-1} \arrow[ur, dash] & & b_m .
\end{tikzcd}$$

Consider the common acceptable sequence $s^* = (s_0, \cdots, s_{i-1}, s_j,$

$\cdots , s_{m-1})$, which is of length l, where $l < m = ab - 1$. Since $\omega(a_m) \neq \omega(b_m)$, A and B would give different output sequences to the input sequence s^*, contradicting our hypothesis that A and B are completely equivalent with respect to R_{AB} $(ab - 2)$. We must therefore assume that the ab pairs of states $(a_0 , b_0), \cdots , (a_m , b_m)$ are distinct, and thus include every possible pair of states of A and B.

Now let a_m' and b_m' be states of A and B respectively such that $a_m' \neq a_m$ and $b_m' \neq b_m$. Then we assert $\omega(a_m) \neq \omega(a_m')$ and $\omega(b_m) \neq \omega(b_m')$. For, if $\omega(a_m) = \omega(a_m')$, then $\omega(a_m') \neq \omega(b_m)$. This is impossible, however, since the pair (a_m' , b_m) is one of the ab distinct pairs of states. The same argument shows $\omega(b_m) \neq \omega(b_m')$. We have now then the inequality $\omega(a_m') \neq \omega(b_m')$. But again this is impossible. This concludes the proof.

Although we cannot say that the bound $ab - 2$ is the best for all pairs (a, b), we will show that $ab - 2$ is very close to the best we can hope for. To do this we will now exhibit a pair of families of finite automata.

Consider first a family of finite antomata $\{A_m\}$, $m \geqq 1$, as follows:

State	Symbol		Output
	0	1	
$^*a_0$	a_1		0
a_1	a_2		0
$\vdots$	$\vdots$		$\vdots$
a_{m-1}	a_m		0
a_m		a_0	0

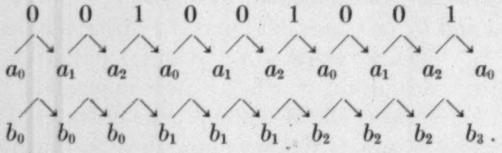
Next, define a family of finite automata $\{B_n\}$, $n \geqq 1$, as follows:

State	Symbol		Output
	0	1	
$^*b_0$	b_0	b_1	0
b_1	b_1	b_2	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
b_{n-1}	b_{n-1}	b_n	0
b_n	b_n		1

For any pair of automata (A_m , B_n), one from each family, the set $R_{A_m B_n}$ of all commonly acceptable sequences consists of all sequences each of which must be of the form

$$\underbrace{0\ 0 \cdots 0}_{m\ 0\text{'s}}\ 1\ \underbrace{0\ 0 \cdots 0}_{m\ 0\text{'s}}\ 1 \cdots \underbrace{0\ 0 \cdots 0}_{m\ 0\text{'s}}\ 1,$$

since these are the only sequences acceptable to A_m. For these two finite automata (A_m, B_n), the minimum length of any input sequence in $R_{A_m B_n}$ that would cause A_m and B_n to give different output sequences would be $ab - \min(a, b)$, where in this case $a = m + 1$ and $b = n + 1$. For instance, the motion diagram for the pair (A_2, B_3) would be


0 0 1 0 0 1 0 0 1
a_0 a_1 a_2 a_0 a_1 a_2 a_0 a_1 a_2 a_0
b_0 b_0 b_0 b_1 b_1 b_1 b_2 b_2 b_2 b_3.

Since b_3 is the only state of B_3 that gives an output symbol of 1, we see that the input sequence (0 0 1 0 0 1 0 0 1) is the first such sequence that causes A_2 and B_3 to give different output symbols.

In general, by the same construction, we find that given two finite automata, one from each of these families, no input sequence of length less than $ab - \min(a, b)$ would enable us to tell them apart. We therefore have

Theorem 4: Theorem 3 would not hold if l were made less than $ab - \min(a, b)$.

In particular, we note that Theorem 3 implies Theorem 4 for the case $\min(a, b) = 2$. For the cases $\min(a, b) > 2$, there may be some slight improvement† possible for Theorem 3.

Actually, Theorem 3 is interesting for another reason. It is essentially a theorem showing the existence of a decision procedure for finding finite automata with a minimum number of states. Historically, the problem of finding finite automata with a minimum number of states was studied and solved in a rather special way. Thus, both Moore[7] and Huffman[8] gave ingenious procedures for state minimization of *complete* finite automata. It was not uncommon for people to assume that these procedures also worked for partial automata before the introduction of several interesting counter-examples by Ginsburg.[12] As we see from Theorem 3, much of the earlier confusion was probably due to a disregard of the idea of acceptable sequences.

VI. FINITE AUTOMATA DEFINED BY INPUT SEQUENCES

Up to now we have shown that finite automata can be described in two different ways. In the definition given in the previous section, a finite automaton is characterized essentially by its state-symbol table. On the other hand, one may characterize a finite automaton by giving its W*-machine program. The latter characterization illustrates the close parallel between computer programming and logical design. In this section, following the earlier work of Kleene,[3] we will consider a third characterization of partial finite automata. This characterization leads to a very interesting algebraic-like structure for finite automata. Our purpose here is to connect this characterization with the others. Much of the work along the approach of Kleene had been pursued and simplified by Myhill[13] and Rabin and Scott.[11] The interested reader may refer to these papers and other unpublished work by Myhill.

---

† In the paper by Rabin and Scott,[11] a theorem similar to Theorem 3 was obtained for the family of complete automata. In view of the fact that they were dealing exclusively with complete automata, their theorem could be considerably improved.

Let A be a finite automaton. A finite input sequence to A is said to be a *signal sequence* for A if this input sequence causes A to terminate in a state whose output is the symbol 1. The set of all signal sequences for a finite automaton A is called the *signal set* for A, and is denoted by $\Gamma(A)$.

Given a finite automaton A, the signal set $\Gamma(A)$ is uniquely defined. On the other hand, if signal sets are to represent finite automata, it would be most desirable that two "different" automata have different signal sets. Let us consider automata A and B given by

A:

State	Symbol		Output
	0	1	
$^*a_0$	a_1		1
a_1		a_1	0

B:

State	Symbol		Output
	0	1	
$^*b_0$		b_2	1
b_1		b_2	0
b_2	b_1		0

If nothing is said about input sequences, one may say that A and B are different, since every input sequence acceptable to A is unacceptable to B and vice versa, although A and B both have the empty set as their signal set. In order to have a clear-cut correspondence between signal sets and finite automata, we must therefore restrict ourselves to acceptable sequences.

Theorem 5: Let A and B be two finite automata and R_{AB} the common acceptable set for A and B. Then A and B are completely equivalent with respect to R_{AB} if and only if A and B have the same signal set.

Proof: From the definition of signal set, it is clear that, if A and B are completely equivalent with respect to R_{AB}, then $\Gamma(A) = \Gamma(B)$. Now suppose A and B have the same signal set but are not completely equivalent with respect to R_{AB}. Then there is some input sequence

$s_0, s_1, \cdots, s_{n-1}$ in R_{AB} giving the motion diagram

$$\begin{tikzcd}
 & s_0 \arrow[dr] & & s_1 \arrow[dr] & & \cdots & & s_{n-1} \arrow[dr] & \\
a_0 \arrow[ur, dash] & & a_1 \arrow[ur, dash] & & a_2 & \cdots & a_{n-1} \arrow[ur, dash] & & a_n \\
 & {} \arrow[dr] & & {} \arrow[dr] & & & & {} \arrow[dr] & \\
b_0 \arrow[ur, dash] & & b_1 \arrow[ur, dash] & & b_2 & \cdots & b_{n-1} \arrow[ur, dash] & & b_n
\end{tikzcd}$$

such that $\omega(a_n) \neq \omega(b_n)$; that is, the output symbols associated with states a_n and b_n are different. Since we are considering only two-symbol automata, it is clear that the input sequence $s_0, s_1, \cdots, s_{n-1}$ cannot be a signal sequence for both A and B. The proof now follows from this contradiction.

We see from this that signal sets indeed represent finite automata. In many ways this is a rather natural characterization. For example, consider a sequential lock on a vault. The vault can be opened only if a given sequence s of symbols is applied to the lock. Any other sequence of input symbols may cause the lock to go into an alarm state. In this case, we may consider the lock as a finite automaton defined by the one-element signal set $\{s\}$.

There are other situations, however, where it seems simpler to describe a finite automaton by its W*-machine program or its state-symbol table. It is therefore not clear in general how a finite automaton is best characterized; as far as we can tell, a great deal depends on personal taste. The next best thing one can do, therefore, is to find ways to go from one form of characterization to another.

We will begin by redefining several operations on sets of finite sequences due to Kleene. Let X and Y be two sets of finite sequences; $X \vee Y$ is then the set *union* of X with Y. By XY, called the *string product* of X with Y, we mean the set of all concatenated finite sequences of the form xy with $x \in X$ $y \in Y$. Finally, by the *closure* of the set X, denoted by X^*, is meant the set

$$X^* = \emptyset \vee X \vee XX \vee XXX \vee \cdots,$$

where $\emptyset$ is the empty set.

To illustrate the use of these operations, let us consider the following automaton A:

State	Symbol		Output
	0	1	
$^*a_0$	a_1	a_0	0
a_1		a_0	1

The signal set of A is then given by

$$\Gamma(A) = 1^* \, 0 \, (1 \, 1^* \, 0)^*,$$

where we have used the notations 0 and 1 to stand for the one-element sets $\{0\}$ and $\{1\}$.

In general, to find the signal set for some finite automaton A is not as straightforward as this example indicates. We will describe below one such procedure.†

Let A be a finite automaton with k states $a_0, a_1, \cdots, a_{k-1}$. Then by $P(a_i, a_j)$ we mean the set of all finite sequences such that beginning with state a_i, each of these sequences causes A to terminate in state a_j. Furthermore, let us denote by $P(a_i, a_j; a'_j)$ the set of all finite sequences such that, beginning with state a_i, each of these sequences not only causes A to terminate in state a_j, but also never causes A to pass through state a_j. In other words, it is permissible for $a_i = a_j$, but in the chain of states $a_i, \cdots, a_j$, the state a_j must not appear other than at either end. Then it is clear that

Lemma 2: $P(a_i, a_j) = P(a_i, a_j; a'_j) \, [P(a_j, a_j; a'_j)]^*$.

More generally, let $s_0, s_1, \cdots, s_{n-1}$ be a sequence in $P(a_i, a_j)$ with the motion diagram

$$\begin{tikzcd}
 & s_0 \arrow[dr] & & s_1 \arrow[dr] & & \cdots & & s_{n-1} \arrow[dr] & \\
a_i \arrow[ur, dash] & & a_{j_1} \arrow[ur, dash] & & a_{j_2} & \cdots & a_{j_{n-1}} \arrow[ur, dash] & & a_j .
\end{tikzcd}$$

We denote by $P(a_i, a_j; a'_{i_1}, a'_{i_2}, \cdots, a'_{i_m})$ a subset of $P(a_i, a_j)$ such that a sequence $s_0, s_1, \cdots, s_{n-1}$ is in $P(a_i, a_j; a'_{i_1}, a'_{i_2}, \cdots, a'_{i_m})$ if and only if the two sets of states $(a_{i_1}, a_{i_2}, \cdots, a_{i_m})$ and $(a_{j_1}, a_{j_2}, \cdots, a_{j_{n-1}})$ are disjoint. In other words, $P(a_i, a_j; a'_{i_1}, a'_{i_2}, \cdots, a'_{i_m})$ is the set of finite sequences such that, beginning with state a_i, each of these sequences not only causes A to terminate in state a_j but also causes A never to go through states $a_{i_1}, a_{i_2}, \cdots, a_{i_m}$. It is permissible, however, for a_i or a_j to be one of the states $a_{i_1}, a_{i_2}, \cdots, a_{i_m}$.

Lemma 3: Let A be a finite automaton with k states $a_0, a_1, \cdots, a_{k-1}$. Then, for all pairs of states a_i, a_j, and for all m, $1 \leqq m \leqq k - 1$,

$$\begin{aligned}
&P(a_i, a_j; a'_{i_1}, a'_{i_2}, \cdots, a'_{i_m}) = \\
&\quad P(a_i, a_j; a'_{i_1}, \cdots, a'_{i_{m+1}}) \vee P(a_i, a_{i_{m+1}}; a'_{i_1}, \cdots, a'_{i_{m+1}}) \\
&\quad [P(a_{i_{m+1}}, a_{i_{m+1}}; a'_{i_1}, \cdots, a'_{i_{m+1}})]^* \, P(a_{i_{m+1}}, a_j; a'_{i_1}, \cdots, a'_{i_{m+1}}).
\end{aligned}$$

Proof: Suppose that an input sequence belongs to the set on the left-hand side. Then this sequence causes A to either go through state $a_{i_{m+1}}$

---

† In an unpublished report shown to me by H. Wang, I found a similar result worked out independently by R. McNaughton and H. Yamada.

or it does not. If it does not, then it clearly belongs to $P(a_i, a_j; a'_{i_1}, \cdots, a'_{i_{m+1}})$. If it does, then it belongs to the second set on the right hand side. Conversely, suppose a sequence belongs to the set on the right hand side. Then it clearly belongs to $P(a_i, a_j; a'_{i_1}, \cdots, a'_{i_m})$, and the proof follows.

Combining the two lemmas, we get

Theorem 6: Let A be a finite automaton with k states $a_0, a_1, \cdots, a_{k-1}$. Let a_0 be the initial state of A and $a_{r_1}, a_{r_2}, \cdots, a_{r_n}$ be all those states of A whose output symbol is one. Then the signal set for A is the union

$$\Gamma(A) = \bigvee_{i=1}^{n} P(a_0, a_{r_i}; a'_{r_i}) [P(a_{r_i}, a_{r_i}; a'_{r_i})]^*,$$

which can be obtained by repeated application of Lemma 3.

As an illustration, let us consider the automaton A below:

State	Symbol		Output
	0	1	
a_0	a_0	a_1	1
a_1	a_1	a_2	1
a_2	a_2	a_1	0

By Lemma 2, we have

$$\Gamma(A) = P(a_0, a_0; a'_0) [P(a_0, a_0; a'_0)]^* \vee P(a_0, a_1; a'_1) [P(a_1, a_1; a'_1)]^*.$$

Now

$$P(a_0, a_0; a'_0) = 0,$$

$$P(a_0, a_1; a'_1) = 0^*\, 1,$$

$$P(a_1, a_1; a'_1) = 0 \vee 1\, 0^*\, 1.$$

Therefore,

$$\Gamma(A) = 0\, 0^* \vee 0^*\, 1\, (0 \vee 1\, 0^*\, 1)^*.$$

The expressions for signal sets can get very lengthy. The problem of reducing the length of these expressions without recourse to an exhaustive search appears very difficult and intriguing.

The next problem we will consider is how to give a state-symbol characterization of signal sets. The procedure we will describe here is a modification of the abstract ideas of Rabin and Scott[11] and Myhill.[13]

Let us begin this discussion of several examples. Let A and B be the following finite automata:

A:

State	Symbol		Output
	0	1	
$^*a_0$	a_1	a_0	0
a_1		a_0	1

B:

State	Symbol		Output
	0	1	
$^*b_0$	b_1	b_1	0
b_1	b_0	b_0	1

with signal sets $\Gamma(A) = 1^* \, 0 \, (1 \, 1^* \, 0)^*$ and $\Gamma(B) = (0 \vee 1) \, [(0 \vee 1) \, (0 \vee 1)]^*$.

Example 1. Suppose we wish to construct an automaton C such that $\Gamma(C) = \Gamma(A) \vee \Gamma(B)$. We begin by defining a set of new states (a_0, b_0), (a_0, b_1), (a_1, b_0), (a_1, b_1), some of which may turn out to be superfluous. The state (a_0, b_0) is defined to be the initial state of C. Beginning with the state (a_0, b_0), we can construct a part of C:

State	Symbol		Output
	1	0	
(a_0, b_0)	(a_1, b_1)	(a_0, b_1)	$\omega(a_0) \vee \omega(b_0) = 0$

where, if we let M_C denote the function taking state-symbol pairs to states for the automaton C and ω be the function taking states to output symbols, then

$$M_C[(a_0, b_0), 0] = (M_A(a_0, 0), M_B(b_0, 0)),$$

$$M_C[(a_0, b_0), 1] = (M_A(a_0, 1), M_B(b_0, 1))$$

and

$$\omega[(a_0, b_0)] = \omega(a_0) \vee \omega(b_0).$$

In this process, we reached two new states (a_1, b_1) and (a_0, b_1). Continuing the process, we eventually get for C the state-symbol table

State	Symbol		Output
	0	1	
$^*(a_0, b_0)$	(a_1, b_1)	(a_0, b_1)	0
(a_0, b_1)	(a_1, b_0)	(a_0, b_0)	1
(a_1, b_1)	b_0	(a_0, b_0)	1
(a_1, b_0)	b_1	(a_0, b_1)	1
b_0	b_1	b_1	0
b_1	b_0	b_0	1

Let us suppose that an input sequence $s_0, s_1, \cdots, s_{n-1}$ belongs to $\Gamma(A)$ and gives the motion diagram

$$\begin{tikzcd}
& s_0 \arrow[dr] & & s_1 \arrow[dr] & \cdots & & s_{n-1} \arrow[dr] & \\
a_0 \arrow[ur, dash] & & a_{i_1} \arrow[ur, dash] & & a_{i_2} & \cdots & a_{i_{n-1}} \arrow[ur, dash] & & a_{i_n}.
\end{tikzcd}$$

Then the same sequence would give rise to a chain of states of C such that the terminal state of this chain must be (a_{i_n}, b_j) for some state b_j of B. Since

$$\omega(a_{i_n}, b_j) = \omega(a_{i_n}) \vee \omega(b_j) = 1,$$

it follows that this input sequence belongs to $\Gamma(C)$ and $\Gamma(A) \subset \Gamma(C)$. In the same way, we may show that $\Gamma(B) \subset \Gamma(C)$.

Conversely, if a sequence in $\Gamma(C)$ gives a chain of states of C: (a_0, b_0), $(a_{i_1}, b_{i_1}), \cdots, (a_{i_n}, b_{i_n})$, then either $\omega(a_{i_n}) = 1$ or $\omega(b_{i_n}) = 1$. Therefore, this input sequence is either in $\Gamma(A)$ or $\Gamma(B)$, and thus $\Gamma(A) \vee \Gamma(B) = \Gamma(C)$.

Example 2. We wish to construct a finite automaton C such that $\Gamma(C) = \Gamma(A)\Gamma(B)$. We begin with the initial state of A as the initial state for C. Now, whenever a state of A is reached whose output symbol is a 1, we must then allow C the opportunity to imitate the behavior of B. In such cases, therefore, new states may be created. Thus, a part of the state-symbol table for C would be

State	Symbol		Output
	0	1	
$^*a_0$	a_1	a_0	0
a_1	b_1	(a_0, b_1)	1

The state (a_0, b_1) is defined by $(M_A(a_1, 1), M_B(b_0, 1))$. In this way, the new state allows C to imitate immediately the behavior of state b_0 of B. Also, if either $\omega(a_0) = 1$ or $\omega(b_1) = 1$, then $\omega(a_0, b_1) = 1$. We may therefore continue this process to get for C the state-symbol table:

State	Symbol		Output
	0	1	
*a_0	a_1	a_0	0
a_1	b_1	(a_0, b_1)	1
b_1	b_0	b_0	0
b_0	b_1	b_1	1
(a_0, b_1)	(a_1, b_0)	(a_0, b_0)	1
(a_1, b_0)	b_1	(a_0, b_1)	1
(a_0, b_0)	(a_1, b_1)	(a_0, b_1)	0
(a_1, b_1)	(b_1, b_0)	(a_0, b_0, b_1)	1
(b_1, b_0)	(b_0, b_1)	(b_0, b_1)	1
(a_0, b_0, b_1)	(a_1, b_0, b_1)	(a_0, b_0, b_1)	1
(a_1, b_0, b_1)	(b_1, b_0)	(a_0, b_0, b_1)	1

where we see that $(a_1, b_0, b_1) = (a_1, b_1)$.

The process can be formulated as follows: If $(a_i, \cdots, a_j, b_k, \cdots, b_m)$ is a new state, then

$$M_C[(a_i, \cdots, a_j, b_k, \cdots, b_m), x]$$
$$= (M_C(a_i, x), \cdots, M_C(a_j, x), M_B(b_k, x), \cdots, M_B(b_m, x))$$

where x is either 0 or 1, and $\omega[(a_i, \cdots, a_j, b_k, \cdots, b_m)] = \omega(a_i) \vee \cdots \vee \omega(b_k) \vee \cdots \vee \omega(b_m)$. Also, if a_i is any state such that $\omega(a_i) = 1$, then $M_C(a_i, x) = (M_A(a_i, x), M_B(b_0, x))$. For all other a_j and for all b_k we have

$$M_C(a_j, x) = M_A(a_j, x),$$
$$M_C(b_k, x) = M_B(b_k, x),$$

where x is again either 0 or 1.

Example 3. We wish to construct an automaton C such that $\Gamma(C) = [\Gamma(A)]^*$. The idea here is that whenever a state of A is reached whose output symbol is a 1, we must allow C the opportunity to begin again at state a_0 of A. Furthermore, since the empty sequence is a member of $\Gamma(C)$, it is necessary to define for C a new initial state C_0 whose output symbol is 1. Following this line of thought, we see that the state-symbol table for C is

State	Symbol		Output
	0	1	
*c_0	a_0	a_0	1
a_0	a_1	a_0	0
a_1	a_1	a_0	1

In general, the process is formulated as follows. If $(a_i, \cdots, a_j)$ is a new state, then

$$M_C[(a_i, \cdots, a_j), x] = (M_C(a_i, x), \cdots, M_C(a_j, x)),$$

where x is either 0 or 1 and

$$\omega[(a_i, \cdots, a_j)] = \omega(a_i) \vee \cdots \vee \omega(a_j).$$

If a_i is any state of A whose output symbol is 1,

$$M_C(a_i, x) = (M_A(a_i, x), M_A(a_0, x)), \qquad x = 0 \text{ or } 1.$$

For all other states a_j of A,

$$M_C(a_j, x) = M_A(a_j, x), \qquad x = 0 \text{ or } 1.$$

The ideas of conversion from signal sets to state-symbol table for a finite automaton are all contained in these examples. Since to state a theorem means a repetition of what we outlined in the examples, we will content ourselves with the following form of Kleene's result.[3]

Remark. Let $\Gamma(A)$ be a set of finite sequences built up from the operations union, string product and closure operating on a finite set of finite sequences. That is, $\Gamma(A)$ is given by a finite expression involving the operations union, string product and closure. Then, following the procedures outlined in Examples 1, 2 and 3, a finite automaton can be constructed having $\Gamma(A)$ as its signal set.

This remark, together with Theorem 6, thus provides the two-way linkage between finite automata and signal sets.

VII. CONCLUDING REMARKS

We have discussed three approaches to a theory of automata and finite automata: the state-symbol table model, the W-machine program model and the signal-set model. Of these, we are most intrigued by the programming model. This approach not only resembles strongly computer programming, but it also offers possibilities of symbol operation and other combinatorial programs, all based on a very simple and elegant program structure. (One other model not studied here is a system proposed by Post.) It is quite possible a combination of these systems may offer deeper insight into the global structure of programming and automata which is lacking at present.

VIII. ACKNOWLEDGMENT

The writer is indebted to E. F. Moore and T. H. Crowley of Bell Telephone Laboratories for their suggestions which led to an improved version of Theorem 4.

REFERENCES

1. Turing, A. M., On Computable Numbers, with an Application to the Entscheidungs Problem, Proc. London Math. Soc., **24,** 1936, p. 230.
2. Wang, H., A Variant to Turing's Theory of Computing Machines, Jour. A.C.M., **4,** 1957, p. 63.
3. Kleene, S. C., Representation of Events in Nerve Nets and Finite Automata, *Automata Studies*, Annals of Math. Studies, No. 34, Princeton Univ. Press, Princeton, N. J., 1956, p. 3.
4. Moore, E. F., A Simplified Universal Turing Machine, Proc. A.C.M. (Sept. 8, 1952), 1953.
5. Ikeno, N., An Example of Universal Turing Machine, (in Japanese), Proc. Inst. Elec. Comm. of Japan, July 1958.
6. Shannon, C. E., A Two-State Universal Turing Machine, *Automata Studies* Annals of Math. Studies, No. 34, Princeton Univ. Press., Princeton, N. J., 1956, p. 157.
7. Moore, E. F., Gedanken Experiments on Sequential Machines, *Automata Studies*, Annals of Math. Studies, No. 34, Princeton Univ. Press, Princeton, N. J., 1956, p. 129.
8. Huffman, D. A., The Synthesis of Sequential Switching Circuits, J. Frank. Inst., **257,** 1954, pp. 161; 275.
9. Mealy, G. H., A Method for Synthesizing Sequential Circuits, B.S.T.J., **34,** 1955, p. 1045.
10. Cadden, W. J., Equivalent Sequential Circuits, I.R.E. Trans., **CT-6,** 1959, p. 30.
11. Rabin, M. O. and Scott, D., Finite Automata and Their Decision Problems, I.B.M. J. Res. & Dev., **3,** 1959, p. 114.
12. Ginsburg, S., On the Reduction of Superfluous States in a Sequential Machine, Jour. A.C.M., **6,** 1959, p. 259.
13. Myhill, J., Finite Automata and Representation of Events, W.A.D.C. Report, 1957.

Transition Probabilities for Telephone Traffic*

By V. E. BENEŠ

(Manuscript received April 21, 1960)

A stochastic model for the occupancy $N(t)$ of a telephone trunk group is specified by the conditions that arriving calls form a renewal process, that holding times have a negative exponential distribution, and that lost calls are cleared. The transition probabilities of $N(t)$ are determined, and their limits are studied. These transition probabilities have practical value in making theoretical estimates of sampling error in traffic measurements, and in the study of overflow traffic.

I. INTRODUCTION

We shall study a stochastic process $\{N(t), t \geqq 0\}$, which is a mathematical model for the occupancy of N service facilities, with no provisions for delays. For example, $N(t)$ can be interpreted as the number of (fully accessible) telephone channels (trunks) out of a group of N such in use at time t, with lost calls cleared. Also, we can think of $N(t)$ as the number of items on order at time t in an idealized inventory situation in which at most N items can be on order at one time (see Arrow, Karlin and Scarf[1]). Throughout the paper we use terminology appropriate to an application to telephone trunking. The process $N(t)$ is determined by the following assumptions:

i. Holding times of trunks are independent, each with the same negative exponential distribution function, of mean γ^{-1}, γ being the "hang-up rate."

ii. Times between successive attempts to place a call (interarrival times) are independent; each has the distribution function $A(\cdot)$, where $A(\cdot)$ is arbitrary except for the condition $A(0) = 0$. This assumption covers Poisson arrivals as a special case. The mean of $A(\cdot)$, when it exists, is denoted by μ_1.

---

* This work was completed while the author was visiting lecturer at Dartmouth College, 1959–60.

iii. There are N trunks, N being finite.

iv. Calls that find all N trunks busy are lost, and are cleared from the system without effect on the flow of arrivals (no retrials). These or similar assumptions appear in Palm,[2] and in Pollaczek;[3,4,5,6] certain properties of $N(t)$ itself have been studied by Takács,[7,8,9] Cohen[10] and Beneš.[11]

II. SUMMARY

The random process of interest is $N(t)$, which is interpreted as the number of trunks in use, or the number of calls in progress, at time t; $N(t)$ is a random step function fluctuating in unit steps from 0 to N. For the most part, we restrict attention to that version of $N(t)$, written $N(t - 0)$, that is continuous from the left.

The present paper is chiefly theoretical in character. It provides (a) formulas for the Laplace transforms of the transition probabilities of the stochastic process $N(t - 0)$, and (b) a statistical description of the calls that *overflow* a trunk group of the kind described in Section I. The formulas will be exemplified and used in a second paper,[12] where specific applications to switch counting and traffic averaging are described.

We begin Section III with a general account telling what transition probabilities are and why they are useful and interesting in traffic theory. The primary result, Theorem 1, can then be stated; it completely characterizes the transition probabilities

$$\Pr\{N(t - 0) = n \mid N(0+) = m\}$$

as functions of t by determining their Laplace transforms, under the restriction that $A(\cdot)$ has a probability density. Section III ends with a computation of some important transition probabilities for Poisson arrivals; practical consequences of these results will be developed in the second paper.[12]

We prove Theorem 1 in the Appendix A. If $y(t)$ is the time elapsed since the last call arrival prior to t, the process $\{N(t - 0), y(t)\}$ is Markov, and we calculate its distributions from the usual Kolmogorov equations. The stationary distribution of this Markov process is determined in Appendix B.

In Appendix C the process $N(t - 0)$ is studied directly in terms of renewal theory and regenerative processes, using results of Smith.[13] No assumptions of absolute continuity are made. This procedure leads to an extension of Theorem 1, and other results outlined in the next paragraphs. (Details are omitted.)

Let R_n be this event: a call arrives and finds n trunks in use. Each occurrence of R_n, where $0 \leqq n \leqq N$, is a regeneration point of $N(t - 0)$, in the sense that the history of $N(t - 0)$ prior to the given occurrence of R_n is statistically irrelevant to the development of $N(t - 0)$ after the occurrence. Let $x_{m,n}$ be the time elapsing from an occurrence of R_m to the next occurrence of R_n. We prove $x_{m,n} < \infty$ with probability one, and, if

$$\mu_1 = \int_0^\infty x\,dA(x) < \infty,$$

then $E\{x_{m,n}\} < \infty$.

The underlying probability functions that we calculate in Appendix C are, for $0 \leqq n \leqq N$:

$$\begin{aligned} Q_n(t) &= \sum_{k=1}^{\infty} \Pr\{k\text{th call arrives before } t \text{ and finds } n \text{ trunks busy}\} \\ &= E\{\text{number of occurrences of } R_n \text{ in } [0,t)\}. \end{aligned} \tag{1}$$

From this interpretation it is apparent that the $Q_n(\cdot)$ are unbounded monotone functions; one may expect them to be ultimately linear. The transition probabilities of $N(t - 0)$ can be represented in terms of the functions $Q_n(\cdot)$ and the transition probabilities of the simple death process with death rate γ per head of population, if the $Q_n(\cdot)$ are evaluated for appropriate initial conditions. This is done in Appendix C. With this representation we investigate the existence of

$$\lim_{t\to\infty} \Pr\{N(t - 0) = n\}.$$

From Theorem 4 and the solutions for the Laplace-Stieltjes transforms of the $Q_n(\cdot)$, this limit, when it exists, can be evaluated explicitly, using the relation

$$E\{x_{n,n}\} = \frac{\mu_1}{p_n},$$

where p_n is the equilibrium probability that an arriving cell will find n trunks in use. (For p_n see Refs. 7 and 11.)

III. TRANSITION PROBABILITIES OF $N(t)$

The *transition probabilities* of a stochastic process x_t tell how likely it is that the random function $x_{(\cdot)}$ take on a value z at a time t, if it is known that it took on the value y at time s. Such a transition probability is written

$$\Pr\{x_t = z \mid x_s = y\}, \tag{2}$$

the vertical bar being read and interpreted as "given that" or "if."

In other words, (2) expresses the relevance of the information that the event $\{x_s = y\}$ has occurred to the likelihood that the event $\{x_t = z\}$ will occur. In still other words, (2) expresses the *dependence* of the event $\{x_t = z\}$ on $\{x_s = y\}$.

The chief practical use of transition probabilities for models of telephone traffic is in computation of covariance functions; these, in turn, are used to compute theoretical estimates of sampling error in actual traffic measurements, such as time averages and switch counts. To see how this happens in a particular case, we consider the use of the continuous time average

$$M(T) = \frac{1}{T}\int_0^T N(t)\,dt$$

as an estimate of the carried load. The variance of M is

$$E\{M^2\} - E^2\{M\} = T^{-2}\int_0^T\int_0^T [E\{N(t)N(s)\} - E\{N(t)\}E\{N(s)\}]ds\,dt. \tag{3}$$

The integrand is just the covariance $R(t,s)$ between $N(t)$ and $N(s)$; if $N(t)$ is stationary in the wide sense, so that $R(t,s) = R(t - s)$, then (3) reduces (by partial integrations) to

$$\text{Var}\,\{M\} = 2T^{-2}\int_0^T (T - t)R(t)\,dt. \tag{4}$$

The covariance $R(t)$ can be written in terms of the transition probabilities of $N(\cdot)$ as

$$R(t) = \sum_{m=0}^{N}\sum_{n=0}^{N} mnp_m \Pr\{N(t) = n \mid N(0) = m\} - \left(\sum_{m=0}^{N} mp_m\right)^2, \tag{5}$$

where $\{p_m\}$ is the stationary distribution of $N(\cdot)$. Formulas (4) and (5) then indicate how the transition probabilities can be used to find the variance of M.

Our basic result concerning transition probabilities is most easily explained and understood after some of the notions used in stating it are discussed. The first few are merely abbreviations; we let

$$\begin{aligned}
A^*(s) &= \int_0^\infty e^{-st}dA(t) = a_0(s),\\
a_n(s) &= A^*(s + n\gamma),\\
X_0 &= 1,\\
X_n &= \frac{1 - a_n(s)}{a_n(s)}X_{n-1} = \prod_{j=1}^{n}\frac{1 - a_j(s)}{a_j(s)}.
\end{aligned}$$

These Laplace transforms enter because we shall be characterizing the Laplace transforms of the transition probabilities in terms of the hang-up rate γ and the transform $A^*(s)$ of the interarrival probability density.

In the summary we have denoted by R_n the event: a call arrives and finds n trunks in use. We let $q_n(t,0)$ be the "density" of R_n at time t, that is, the rate at which R_n is occurring at t, and we let

$$b_n(t) = \sum_{j=n}^{N} \binom{j}{n} q_j(t,0) \tag{6}$$

be the associated binomial moment. From (1), it can be seen that $dQ_n/dt = q_n(t,0)$, when the former exists. The $b_n(\cdot)$ and the $q_n(\cdot,0)$ are also related by the inversion formula

$$q_n(t,0) = \sum_{j=0}^{N-n} (-1)^j \binom{n+j}{n} b_j(t).$$

More generally, we use $q_n(t,u)$ as a density function in the variable u with the heuristic meaning

$$q_n(t,u)du = \Pr\{N(t-0) = n \quad \text{and} \quad u < y(t) \leqq u + du\}.$$

We can now state

Theorem 1: The transition probabilities of $N(t-0)$ may be determined from the generating function formula

$$\begin{aligned} E\{x^{N(t-0)}\} = {} & \int_0^t \sum_n q_n(t-y,0)[P_y(x)]^{n+1-\delta_{Nn}}[1-A(y)]\,dy \\ & + \int_t^\infty \sum_n q_n(0,y-t)[P_t(x)]^n \frac{1-A(y)}{1-A(y-t)}\,dy, \end{aligned} \tag{7}$$

where

$$P_u(x) = 1 + (x-1)e^{-\gamma u},$$

and the Laplace transforms of the binomial moments $b_n(\cdot)$ are given by

$$b_n^*(s) = (X_n)^{-1}\left\{b_0^* - \sum_{j=1}^{n}\left[\binom{N}{j-1} b_N^* - \frac{k_n^*}{a_n(s)}\right] X_{j-1}\right\},$$

$$b_0^*(s) = \frac{k_0^*}{1-A^*(s)},$$

$$b_N^*(s) = \frac{\dfrac{k_0^*}{1-a_0(s)} + \displaystyle\sum_{j=1}^{N} \frac{X_{j-1}k_j^*(s)}{a_j(s)}}{\displaystyle\sum_{n=0}^{N}\binom{N}{n} X_n},$$

where

$$k_n^* = \text{Laplace transform of} \int_0^\infty \sum_{j=n}^{N} \binom{j}{n} q_j(0,u)\, e^{-n\gamma t} \frac{a(t+u)\, du}{1 - A(u)} .$$

The k_n^* introduce dependence on the boundary conditions at $t = 0$ expressed by the functions $q_n(0,u)$. The Kronecker symbol δ_{Nn} in (7) indicates that a call is lost if it finds all trunks busy.

To show how Theorem 1 can be used we shall compute the Laplace transforms of

$$\Pr\{N(t) = N \mid N(0) = m\}, \qquad m = 0, 1, \cdots, N,$$

in the important special case of Poisson arrivals at rate a, for which a great simplification of the formulas occurs. In this case,

$$A(t) = \begin{cases} 1 - e^{-at}, & t \geqq 0, \\ 0 \quad , & t < 0. \end{cases}$$

Also, we set γ = hang-up rate = 1, which amounts to measuring time in units of mean holding time; then

$$a_n(s) = \frac{a}{a + s + n}. \tag{8}$$

Our choice of the transition probability to the "all trunks busy" condition $\{N(t) = N\}$ as an example is not arbitrary; it turns out that, in many cases, including Poisson arrivals, the mean of $N(t)$ and the covariance depend only on the transition probability to the "boundary" condition $\{N(t) = N\}$. A similar situation occurs in the theory of queues with one server: the mean delay can be written as an integral of the probability of being on the "boundary," i.e., the chance that the server is idle.[14]

Since arrivals are Poisson, the $y(\cdot)$ process is in fact superfluous, and we may assume $N(0) = m, y(0) = 0$, so that

$$\begin{aligned} k_n^*(s) &= \int_0^\infty e^{-st-nt-at} a \binom{m}{n} dt = \binom{m}{n} \frac{a}{a + s + n}, & n \leqq m, \\ &= 0, & n > m. \end{aligned} \tag{9}$$

In formula (7) (Theorem 1) set $x = 1 + w$, and take Laplace transforms with respect to t; the coefficient of w^N is

$$\int_0^\infty e^{-st} \Pr\{N(t) = N \mid N(0) = m\}\, dt$$

$$= \int_C^\infty e^{-st-Nt-at}\, dt[q_N^* + q_{N-1}^* + \delta_{Nm}],$$

where $q_n^*(s)$ is the transform of $q_n(t,0)$. Now, from (6), (9) and (24) we find

$$q_N^* + q_{N-1}^* = \frac{q_N^* - \delta_{Nm}a_N(s)}{a_N(s)};$$

hence

$$\int_0^\infty e^{-st} \Pr\{N(t) = N \mid N(0) = m\}\, dt = \frac{q_N^*}{a}. \tag{10}$$

This result can also be obtained heuristically as follows:

$$\begin{aligned} a &= \text{total density of arrivals at } t \\ &= q_N(t,0) + a[1 - \Pr\{N(t-0) = N\}]; \end{aligned}$$

taking Laplace transforms, we get (10).

From Theorem 1 and (10), we find

$$\frac{q_N^*}{a} =$$

$$a^{-1}\frac{\binom{m}{0} + \binom{m}{1}\frac{1-a_0(s)}{a_0(s)} + \cdots + \binom{m}{m}\frac{[1-a_0(s)]\cdots[1-a_{m-1}(s)]}{a_0(s)\cdots a_{m-1}(s)}}{\frac{1-a_0(s)}{a_0(s)}\sum_{n=0}^{N}\binom{N}{n}X_n}.$$

But, for our example, (8) implies

$$\frac{1-a_n(s)}{a_n(s)} = \frac{s+n}{a};$$

hence, defining (after J. Riordan in the Appendix to Wilkinson[15]) the "sigma" functions $\sigma_k(m)$ by

$$\sigma_0(m) = \frac{a^m}{m!}, \qquad \sigma_k(m) = \sum_{j=0}^{m}\binom{k+j-1}{j}\frac{a^{m-j}}{(m-j)!},\dagger$$

with m (but not k) an integer, we can show that $a^{-1}q_N^*$ reduces to

$$\int_0^\infty e^{-st} \Pr\{N(t) = N \mid N(0) = m\}\, dt = \frac{a^{N-m}n!\sigma_s(m)}{N!s\sigma_{s+1}(N)}. \tag{11}$$

This and similar results for Poisson arrivals have been found by S. O. Rice in unpublished work.

---

† The "sigma" functions are related to the Poisson-Charlier polynomials $p_n(x) = a^{n/2}(n!)^{\frac{1}{2}}\sum_{j=0}^{n}(-1)^{n-j}\binom{n}{j}j!a^{-j}\binom{x}{j}$ by $\sigma_k(m) = (-a^{\frac{1}{2}})^m(m!)^{-\frac{1}{2}}p_m(-k)$. See Szegö.[16]

Since the event $\{N(t) = N\}$ (the "all trunks busy" condition) is of primary interest, the transition probability

$$p_{NN}(t) = \Pr \{N(t) = N \mid N(0) = N\}$$

has been used (e.g. by Kosten[17]) as a "recovery" or "relaxation" function that is characteristic of the dynamic behavior of the system, especially of its approach to equilibrium from the "all trunks busy" condition. Such a function has been computed from (11) and plotted as Fig. 1, for a (heavy) load of 10 erlangs offered to 5 trunks, giving a loss probability of 0.563.

IV. OVERFLOW TRAFFIC

In the design and engineering of trunking plans in telephony, it is common practice to offer the calls lost by one trunk group to a second or overflow group. It has been discovered that the right choices of group size and the pooling of overflow traffic can lead to efficient trunking arrangements, called *graded multiples*. For this reason, some theoretical work, as well as much empirical study, has been devoted to the statistical behavior of overflowing calls. The principal references are to Brockmeyer,[18] Cohen,[10] Kosten,[19] Palm,[2] Takács,[7,8,9] and Wilkinson.[15]

In accordance with current usage in mathematical literature, let us refer to a sequence of mutually independent, identically distributed, positive random variables as a *renewal process*. The interarrival times that we have assumed in the model describing the trunk group then form a renewal process. It has been shown by Palm[2] that, if calls arriving in

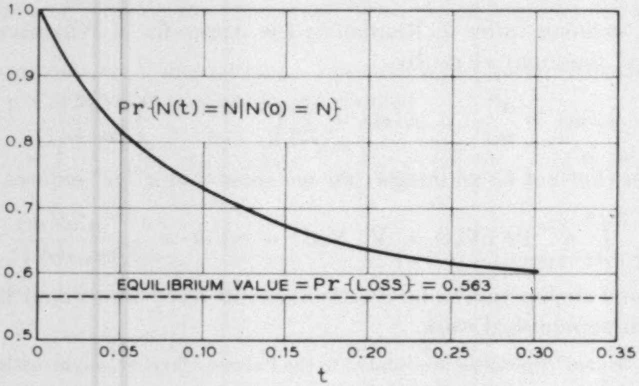


Fig. 1 — "Recovery function" $\Pr \{N(t) = N \mid N(0) = N\}$ for $N = 5$ trunks and $a = 10$ erlangs (heavy traffic).

a renewal process are served by a finite group of trunks, with exponential holding times and lost calls cleared, then the overflowing calls can also be described by a renewal process. That is, the time intervals between successive overflowing calls are mutually independent and identically distributed. Palm also showed how the distribution function of these interoverflow times can be calculated from the interarrival distribution, the hang-up rate and the group size.

We can deduce Palm's results in a simple way from our basic theorem and give a general formula for the Laplace-Stieltjes transform of the interoverflow distribution. Let $O_N(t)$ be the average number of overflows occurring in the closed interval $[0,t]$, assuming that an overflow occurred at time 0. Thus $O_N(t)$ is the particular form of $Q_N(t)$ that arises when $u_1 = 0$ and $N(0-) = N$. We use $G(u)$ to denote the distribution function of the interoverflow times. Since these times are independent, it can be seen that

$$O_N(t) = U(t) + \int_0^t O_N(t-u)dG(u), \qquad t \geqq 0, \tag{12}$$

where $U(t)$ is 1 for $t \geqq 0$, and 0 otherwise. If $O_N^*(s)$ and $G^*(s)$ are the respective Laplace-Stieltjes transforms of O_N and G, then (12) implies

$$G^*(s) = \frac{O_N^*(s) - 1}{O_N^*(s)},$$

which determines $G(u)$ uniquely if O_N^* is known.

Since, as noted, O_N^* is the particular case of Q_N^* arising when $u_1 = 0$ and $N(0-) = N$, a formula for it can be found from (32). In the particular case being considered

$$K_n^* = \binom{N}{n},$$

and so O_N^* is given by

$$\frac{1}{1 - A^*(s)} \cdot \frac{1 + \binom{N}{1}\frac{1 - a_0(s)}{a_1(s)} + \cdots + \binom{N}{N}\frac{[1 - a_0(s)] \cdots [1 - a_{N-1}(s)]}{a_1(s) \cdots a_N(s)}}{\sum_{n=0}^{N}\binom{N}{n}\frac{[1 - a_1(s)] \cdots [1 - a_n(s)]}{a_1(s) \cdots a_n(s)}}. \tag{13}$$

If $\mu_1 = \int_0^\infty x\, dA(x) < \infty$, the mean time between overflows is

$$\frac{\mu_1}{p_N},$$

where p_N is the equilibrium probability of loss (studied in Refs. 7 and 11).

For $N = 1$, (13) gives

$$O_1^* = \frac{1 - A^*(s) + A^*(\gamma + s)}{1 - A^*(s)}, \tag{14}$$

$$G^* = \frac{A^*(\gamma + s)}{1 - A^*(s) + A^*(\gamma + s)}. \tag{15}$$

Since $A^*(\gamma + s)$ is the Laplace-Stieltjes transform of

$$L_1(t) = \int_0^t e^{-\gamma u}\, dA(u),$$

(15) can be inverted to give

$$G(t) = \sum_{n=1}^{\infty} \varphi_n(t) = \{\mathfrak{F}(A)\}(t), \tag{16}$$

where

$$\varphi_1 = L_1,$$
$$\varphi_{n+1} = \varphi_n \bigstar (A - L_1)$$

and "$\bigstar$" denotes Stieltjes convolution.

Formula (13) agrees with the recurrence relation given by Palm[2] for the overflow distribution from N trunks. The "one-trunk" case of (14) through (16) is important theoretically because all other cases can be obtained from it by iteration. Formula (16) defines a mapping $G = \mathfrak{F}(A)$ and the interoverflow distribution for N trunks can be written as $\mathfrak{F}^N(A)$, the Nth iterate.

For one trunk, the first two moments of the interoverflow time u are

$$E\{u\} = \frac{\mu_1}{a_1},$$

$$E\{u^2\} = \frac{\mu^2}{a_1} + \frac{2\mu_1^2}{a_1^2}\left[1 - \frac{1}{\mu_1}\int_0^\infty te^{-\gamma t}\, dA(t)\right],$$

where $\mu_i = \int u^i\, dA$. In particular, the ratio of second to first moment is

$$\frac{E\{u^2\}}{E\{u\}} = \frac{\mu_2}{\mu_1} + \frac{2\mu_1}{a_1}\left[1 - \frac{1}{\mu_1}\int_0^\infty te^{-\gamma t}\, dA(t)\right],$$

so that the mapping $\mathfrak{F}$ always *increases* this ratio, by an amount proportional to $E\{u\}$.

APPENDIX A

Approach Using a Markov Process

Let $N(t)$ be the number of trunks in use at time t. To study the distribution of $N(t)$ we introduce the two-dimensional process $\{N(t),y(t)\}$, where $y(t)$ is the length of the time interval from t back to the last arrival epoch prior to t. We assume that $A(\cdot)$ is absolutely continuous, with a continuous density $a(\cdot)$.

The reason for using the two-dimensional variate is that, unless arrivals are Poisson, the $N(t)$ process by itself is not Markov. To avail ourselves of the functional equations satisfied by the distributions of Markov processes, we include $y(t)$ in the "state of the system." This inclusion does result in a Markov process. The device of "Markovization" by the inclusion of variables has been suggested and developed by Cox,[20] and also has been used by Takács.[7,8,9]

It is natural physically to think of the random functions $N(t)$ and $y(t)$ as being continuous from the right. However, we shall assume only that $y(t)$ is always defined to be equal to $y(t + 0)$, and shall study the two processes $\{N(t + 0),y(t + 0)\}$ and $\{N(t - 0),y(t + 0)\}$ jointly.

That $N(t - 0)$ and $N(t + 0)$ are not the same process is clear: $N(t-0) = N$ and $y(t) = 0$ imply $N(t+0) = N$; but, if $N(t + 0) = N$, $y(t) = 0$, then $N(t - 0) = N$ or $N - 1$ according as the call that just arrived is lost or accommodated. The analysis of $N(t - 0)$ and $N(t + 0)$ shall be carried out in terms of two sets of probability density functions, $p_n(t,y)$ and $q_n(t,y)$, where

$$p_n(t,y)dy = \Pr\{N(t + 0) = n \quad \text{and} \quad y < y(t) \leqq y + dy\},$$

$$q_n(t,y)dy = \Pr\{N(t - 0) = n \quad \text{and} \quad y < y(t) \leqq y + dy\}.$$

Lemma: $p_n(t,y) = q_n(t,y)$ for almost all y.

Proof: Let P be a basic probability measure determined by our assumptions (i) through (iv) of Section I; P is defined for sets of elements ω in a space Ω. We assume further that $N(t,\omega)$ is separable, so that

$$S_\epsilon = \bigcap_{0<u<\epsilon} \{\omega : N(t - u) = N(t + 0) = N(t + u)\}$$

is a measurable set.

Now if $y(t) > \delta > \epsilon$, then $y(t - \epsilon) = y(t) - \epsilon$, almost surely, and

$$\Pr\{S_\epsilon \mid N(t + 0), y(t) > \delta\} \geqq e^{-2\gamma N\epsilon} \frac{1 - A(\delta + \epsilon)}{1 - A(\delta)}$$

independently of $N(t + 0)$, almost everywhere, so that

$$\Pr\{S_\epsilon \mid y(t) > \delta\} \geqq e^{-2\gamma N\epsilon} \frac{1 - A(\delta + \epsilon)}{1 - A(\delta)}.$$

The sets S_ϵ are monotone nondecreasing, so $S_0 = \lim S_\epsilon$ as $\epsilon \to 0$ is measurable, and

$$\Pr\{S_0 \mid y(t) > \delta\} = 1, \qquad \text{almost everywhere } [P], \tag{17}$$

and S_0 is the ω-set on which $N(t)$ is constant in some interval $(t - u, t+u)$. The lemma follows from (17).

It remains to establish the relationship between $p_n(t,y)$ and $q_n(t,y)$ when $y = 0$. From our previous remarks about $N(t - 0)$ and $N(t + 0)$ it can be seen that

$$\begin{aligned} p_N(t,0) &= q_N(t,0) + q_{N-1}(t,0), \\ p_n(t,0) &= q_{n-1}(t,0), \qquad 1 \leqq n \leqq N - 1, \\ p_0(t,0) &= 0. \end{aligned}$$

To formulate the Kolmogorov equations for $p_n(t,\cdot)$ and $q_n(t,\cdot)$, we need the function $\lambda(\cdot)$ defined by

$$\lambda(y) = \frac{a(y)}{1 - A(y)}, \qquad A(y) < 1.$$

This is the probability density that an interarrival time will end in the next dy, given that it has lasted a time y to date. The functions $q_n(t, \cdot)$, where $0 \leqq n \leqq N$, (with $q_{N+1} \equiv 0$), satisfy the difference-differential system

$$\left[\frac{\partial}{\partial t} + \frac{\partial}{\partial y} + \gamma n + \lambda(y)\right] q_n = \gamma(n + 1) q_{n+1}, \tag{18}$$

and the behavior of the densities $q_n(t, \cdot)$ for $y = 0$ is determined by the additional condition

$$q_n(t,0) = \int_0^\infty q_n(t,y)\lambda(y)\, dy. \tag{19}$$

We introduce the generating function

$$\psi(x,t,y) = \sum_{n=0}^{N} x^n q_n(t,y),$$

and from (18) obtain

$$\left[\frac{\partial}{\partial t} + \frac{\partial}{\partial y} + \gamma(x-1)\frac{\partial}{\partial x} + \lambda(y)\right]\psi = 0, \tag{20}$$

whose general solution is

$$\psi(x,t,y) = K\{t-y, e^{-\gamma y}(x-1)\}[1-A(y)].$$

Before continuing, we note that the functions $p_n(t,y)$ also satisfy the system (18), but that the analog of (19) is

$$p_{n+1-\delta_{N,n}}(t,0) = \int_0^\infty p_n(t,y)\lambda(y)\,dy, \tag{21}$$

where the Kronecker δ symbol is used to indicate that an arriving call is lost if it finds all N trunks busy. The generating function $\varphi(x,t,y)$ of the $p_n(t,y)$ is also a solution of (20).

The function $\psi(x,t,\cdot)$ is y-continuous for $y > 0$, so, from the lemma proved previously, we conclude that $\psi(x,t,y) = \varphi(x,t,y)$ almost everywhere in y, and that

$$\lim_{y\to 0}\psi(x,t,y) = \varphi(x,t,0).$$

Because of the "lost calls cleared" assumption, we must have

$$\begin{aligned}\psi(x,t,0+) &= x\psi(x,t,0) - x^N(x-1)q_N(t,0)\\ &= \varphi(x,t,0),\end{aligned}$$

so that ψ is discontinuous in y at $y = 0$.

Let $P_y = P_y(x)$ abbreviate $1 + (x-1)e^{-\gamma y}$. It can be verified that the function $K(\cdot,\cdot)$ in the solution of (20) is given by

$$\begin{aligned}K(u,z) &= (1+z)\psi(1+z,u,0) - z(1+z)^N q_N(u,0), &\quad t \geqq y,\\ &= \frac{\psi(1+ze^{\gamma u},0,-u)}{1-A(-u)}, &\quad t < y,\end{aligned}$$

for the solution $\psi(x,t,y)$. From this we find that, for $t \geqq y$,

$$\psi(x,t,y) =$$

$$P_y\psi(P_y,t-y,0)[1-A(y)] - (P_y-1)P_y^N q_N(t-y,0)[1-A(y)],$$

while, for $t < y$,

$$\psi(x,t,y) = \psi(P_t,0,y-t)\,\frac{1 - A(y)}{1 - A(y-t)}.$$

The solution for $\varphi(x,t,y)$ is analogous; in view of this and of the close relationship between φ and ψ, only ψ shall be treated from now on.

The function $\psi(x,0,y)$ represents initial conditions, and is considered as given. To find $\psi(x,t,0)$, we use the integral condition (19), and conclude that

$$\begin{aligned}\psi(x,t,0) = &\int_0^t \psi(P_y,t-y,0)P_y a(y)\,dy \\ &- \int_0^t (P_y - 1)P_y^N q_N(t-y,0)a(y)\,dy \\ &+ \int_t^\infty \psi(P_t,0,y-t)\,\frac{a(y)\,dy}{1 - A(y-t)}.\end{aligned} \tag{22}$$

To solve the functional-integral equation (22), we set $x = 1 + w$, and equate coefficients of like powers of w. This yields

$$\begin{aligned}b_n(t) = &\int_0^t \Big[b_n(t-y) + b_{n-1}(t-y) \\ &- \binom{N}{n-1} b_N(t-y)\Big] e^{-n\gamma y}a(y)\,dy + k_n(t), \qquad n \geqq 0,\end{aligned} \tag{23}$$

where

$$b_n(t) = \sum_{j=n}^{N} \binom{j}{n} q_j(t,0),$$

$$k_n(t) = \int_0^\infty b_n(u)e^{-n\gamma t}\,\frac{a(t+u)\,du}{1 - A(u)}.$$

Note that

$$b_0(t) = \sum_{n=0}^{N} q_n(t,0) = \psi(1,t,0).$$

Let the Laplace transforms of $b_n(\cdot)$, $k_n(\cdot)$ be $b_n^*(s)$, $k_n^*(s)$, respectively. We obtain a simple recurrence for the b_n^* by applying the Laplace transformation to (23). The recurrence is

$$b_n^* = a_n(s)\left\{b_n^* + b_{n-1}^* - \binom{N}{n-1} b_N^*\right\} + k_n^*, \qquad n \geqq 0, \tag{24}$$

where

$$A^*(s) = \int_0^\infty e^{-su}\, dA(u),$$

$$a_n(s) = A^*(s + n\gamma).\dagger$$

To find b_0^*, let x approach 1 in (22); then $P_y(x)$ goes to 1, and we reach the following renewal equation for $b_0(t)$:

$$b_0(t) = \int_0^t b_0(t - y)a(y)\, dy + \int_0^\infty \frac{\psi(1,0,u)}{1 - A(u)}\, a(t + u)\, du. \quad (25)$$

It can be verified that the last term on the right of (25) is just $k_0(t)$; upon solving (25) by transforms, we find that $b_0^* = k_0^*/[1 - A^*]$.

It can be seen that $b_0(t)$ is the density of arrivals at the time t; thus $b_0(t)$ is a familiar function of renewal theory, for which the reader is referred to Smith[21] and the bibliography therein.

The solution of the recurrence (24) is

$$b_n^* = (X_n)^{-1}\left\{b_0^* - \sum_{j=1}^{n}\left[\binom{N}{j-1} b_N^* - \frac{k_n^*}{a_n(s)}\right] X_{j-1}\right\}, \quad (26)$$

where

$$X_0 = 1,$$

$$X_n = \frac{1 - a_n(s)}{a_n(s)}\, X_{n-1}.$$

In particular, the Laplace transform of the density (at t) of arrivals finding all trunks busy is given by

$$b_N^* = q_N^* = \int_0^\infty e^{-st} q_N(t,0)\, dt = \frac{\dfrac{k_0^*}{1 - a_0(s)} + \displaystyle\sum_{j=1}^{N} \frac{X_{j-1}k_j^*(s)}{a_j(s)}}{\displaystyle\sum_{n=0}^{N} \binom{N}{n} X_n}. \quad (27)$$

The generating function of distr $\{N(t - 0)\}$ is

$$\begin{aligned} E\{x^{N(t-0)}\} &= \int_0^\infty \psi(x, t, y)\, dy \\ &= \int_0^t \sum_n q_n(t - y,0)[P_y(x)]^{n+1-\delta_{Nn}}[1 - A(y)]\, dy \quad (28) \\ &\quad + \int_t^\infty \sum_n q_n(0,y - t)[P_t(x)]^n\, \frac{1 - A(y)}{1 - A(y - t)}\, dy. \end{aligned}$$

---

† The functions $a_n(s)$ are to be distinguished from the constants a_n of Ref. 11, which use the same model and notation. The two quantities are related by $a_n = A^*(n\gamma) = a_n(0)$.

APPENDIX B

The Stationary Distribution of $\{N(t),y(t)\}$

We now consider which initial distributions $q_n(0,u)$ for $\{N(0+), y(0+)\}$ are stationary, i.e., are invariant under the transition probabilities of the Markov process $\{N(t-0),y(t)\}$, studied in Appendix A. Intuitively, since we show in Theorem 3 of Appendix C that a limiting distribution exists as $t \to \infty$, we expect this limit to give the stationary distribution. This is the content of

Theorem 2: If $A(u)$ has a continuous derivative and $\mu_1 < \infty$, the x,u function

$$\sum_n p_n P_u^{\,n+1-\delta_{Nn}}(x)\,\frac{1-A(u)}{\mu_1}, \qquad u \geqq 0, \quad (29)$$

generates the unique stationary distribution of $\{N(0+),y(0)\}$; (29) is a generating function in x and a probability density in u. The number p_n is the equilibrium probability that an arriving customer find n trunks busy.

To show that (29) generates a stationary distribution, it is sufficient to prove that the choice of (29) for the initial condition makes each $q_n(t,0) = p_n/\mu_1$ for all t. This is equivalent to

$$q_n^*(s) = \frac{p_n}{s\mu_1},$$

or to

$$b_n^*(s) = \frac{b_n}{s\mu_1},$$

with

$$b_n = \sum_n^N \binom{j}{n} p_j .$$

In order to use the recurrence (24) and the formula (27) for q_N^*, we must first calculate the quantities k_n^* imposed by (29). Now $k_n(t)$ is the nth binomial moment associated with the generating function

$$\int_t^\infty \psi(P_t,0,y-t)\,\frac{dA(y)}{1-A(y-t)}$$

for $\psi(x,0,u)$ given by (29). Thus, $k_n(t)$ is associated with

$$\mu_1^{-1}\sum_n \int_t^\infty p_n P_{y-t}^{\,1+n-\delta_{Nn}}[P_t(x)]\,dA(y).$$

This is equal to

$$\mu_1^{-1} \sum_n \int_0^\infty p_n P_{u+t}^{1+n-\delta_{Nn}}(x)\, dA(u);$$

and so, for $n > 0$,

$$k_n(t) = \mu_1^{-1} \left\{ b_n + b_{n-1} - \binom{N}{n-1} p_N \right\} \int_t^\infty e^{-n\gamma u}\, dA(u).$$

The Laplace transform of this is

$$k_n^* = \left\{ b_n + b_{n-1} - \binom{N}{n-1} p_N \right\} \frac{a_n - a_n(s)}{s\mu_1}, \qquad n > 0.$$

For $n = 0$,

$$k_0(t) = \frac{1 - A(t)}{\mu_1},$$

$$k_0^* = \frac{1 - A^*(s)}{s\mu_1} = \frac{1 - a_0(s)}{s\mu_1}.$$

We now note that, for these k_n^*, the condition $b_N^* = p_N/s\mu_1$ implies $b_n^* = b_n/s\mu_1$ for all lower n. This can be proved by induction from (24). We now substitute these k_n^* in (27) for q_N^* ($= b_N^*$). If we divide out a factor $[1 - a_0(s)]$ in the numerator, the first term is $1/s\mu_1$; the general term is

$$\left[b_n + b_{n-1} - \binom{N}{n-1} p_N \right] \frac{[1 - a_1(s)] \cdots [1 - a_{n-1}(s)][a_n - a_n(s)]}{(s\mu_1)a_1(s) \cdots a_n(s)}.$$

Using the recurrence of Ref. 11,

$$b_n = a_n \left[b_n + b_{n-1} - \binom{N}{n-1} p_N \right], \qquad n > 0,$$

we find after much algebra that $q_N^* = p_N/s\mu_1$, which proves the theorem. The stationary value p_N/μ_1 for the density $q_N(t,0)$ has the following physical interpretation: $1/\mu_1$ is the equilibrium density of arrivals, and p_N is the chance that such an arrival find all trunks busy.

The uniqueness of the stationary distribution follows from that of the limiting distribution as $t \to \infty$. For two distinct stationary distributions of necessity give rise to distinct limits, contradicting Theorem 4 of Appendix C.

The analog of Theorem 2 for the more general formulation of Appendix C is proved by the same form of argument that established Theorem 2, with the difference that Laplace-Stieltjes transforms are involved, and that special mention must be made of the "periodic" case, in which $A(\cdot)$ is concentrated on a lattice.

APPENDIX C

Approach Via Renewal Theory and Regenerative Processes

This last appendix is a quick sketch of results for general distributions $A(\cdot)$; no proofs are given.

Smith[13] has defined a *regenerative* stochastic process $x(t)$ as one for which there is an event R such that, if R occurs at t, then knowledge of $x(s)$ for $s < t$ loses all prognostic value, and the future development of $x(\tau)$ for $\tau > t$ depends only on the fact that R occurred at t. The points at which R occurs are called *regeneration* points of the process.

Let R_n denote the event: an arriving customer finds n trunks busy. Since the interarrival times form a renewal process, each point in time at which R_n occurs is a regeneration point of $N(t - 0)$, for all $0 \leqq n \leqq N$. In fact, we have already[11] made use of this property of the arrival process in constructing the imbedded Markov chain.

We are therefore in a position to use Smith's results[13] directly. The regenerative property of R_n implies that the time intervals between successive occurrences of R_n form a renewal process, i.e., a sequence of independent, identically distributed variates. To apply the results of Ref. 13 we must investigate whether these variates are proper, i.e., finite almost surely, and whether they have finite expectations. We content ourselves with the following result:

Theorem 3: Let $x_{m,n}$ be the time elapsing from an occurrence of R_m until the next occurrence of R_n. Then

$$x_{m,n} < \infty \text{ with probability 1,}$$

and, if the mean interarrival time $\mu_1 = \int x \, dA < \infty$, then

$$E\{x_{m,n}\} < \infty.$$

We use the following notations:

$$u_i = \text{the } i\text{th interarrival time, } i = 1, 2, 3, \cdots,$$

$$r_i(m) = \text{the time interval between the } (i-1)\text{th and the } i\text{th occurrences of } R_m,$$

$$U_k = \sum_1^k u_i = \text{the epoch of the } k\text{th arrival,}$$

$$X_k(m) = \sum_1^k r_i(m) = \text{the epoch of the } k\text{th occurrence of the event } R_m ,$$

$$T_t = \text{the epoch of the last arrival prior to } t \text{ and after } 0.$$

The u_i all have the common distribution $A(u)$, except for u_1, which has G. Also, the $r_i(m)$ have a common distribution, except for $r_1(m)$.

During the interval (T_t, t), the process $N(x - 0)$ forms a pure death process whose transition probabilities $P_{m,n}(\cdot)$ are known. Let $U(x)$ be the unit step function at 0 and δ_{mN} the Kronecker delta. The probability that $N(t - 0) = n$ can be represented as

$$\Pr\{N(t - 0) = n\} = E\{p_{N(0+),n}(t)U(u_1 - t)\}$$

$$+ \sum_{n-1 \leqq m \leqq N} \int_0^t p_{m+1-\delta_{mN},n}(t - u)\, d_u \Pr\{T_t \leqq u \text{ and } N(T_t - 0) = m\},$$

where the measure implicit in the E operation is the joint distribution of $N(0+)$ and u_1. With the notations just introduced in mind, it can be verified that

$$\Pr\{T_t < u \quad \text{and} \quad N(T_t - 0) = m\} = \sum_{k=1}^{\infty} \Pr\{T_t = X_k(m) \leqq u\}$$

$$= \int_0^u [1 - A(t - v)]\, d_v \sum_k \Pr\{U_k \leqq v \quad \text{and} \quad N(U_k - 0) = m\}$$

$$= \int_0^u [1 - A(t - v)]\, d_v \sum_k \Pr\{X_k(m) \leqq v\},$$

the series being absolutely convergent. By introducing

$$Q_n(t) = \sum_k \Pr\{U_k \leqq t \quad \text{and} \quad N(U_k - 0) = n\},$$

we can write

$$\Pr\{N(t - 0) = n\} = E\{p_{N(0+),n}(t)U(u_1 - t)\}$$

$$+ \sum_{n-1 \leqq m \leqq N} \int_0^t p_{m+1-\delta_{mN},n}(t - v)[1 - A(t - u)]\, dQ_m(v).$$

This representation has been used by Takács[7] to study lim Pr $\{N(t - 0) = m\}$ as $t \to \infty$ by methods similar to those used in the proof of Theorem 4.

We can now describe directly some conditions under which Pr $\{N(t -$

$0) = n\}$ goes to a limiting distribution as $t \to 0$, independent of initial conditions. The result is due essentially to Takács.[7]

Theorem 4: If $A(u)$ is not periodic, if $u_1 < \infty$ almost surely, if $\mu_1 = E\{u_i\} < \infty, i > 1$, then

$$\lim_{t\to\infty} \Pr\{N(t-0) = n\} = \sum_{n-1\leqq m\leqq N} \int_0^\infty p_{n+1-\delta_{mN},n}(u)\, \frac{[1-A(u)]}{E\{y_{m,m}\}}\, du.$$

This result follows at once from the previous results of this section and Theorem 2 of Smith,[13] upon noting that $p_{n,k}(u)$ is a linear combination of monotone decreasing functions. From Theorem 3 of Smith[21] there also follows

Theorem 5: If $A(u)$ has period p, if $u_1 < \infty$ almost surely, if $\mu_1 = E\{u_i\} < \infty, i > 1$, and $0 \leqq y < p$, then

$$\lim_{k\to\infty} \Pr\{N(kp+y-0) = n\} =$$

$$\sum_{n-1\leqq m\leqq N} \sum_{k\geqq 0} p_{m+1-\delta_{mN},n}(kp+y)\, \frac{[1-A(kp+y)]}{E\{x_{m,m}\}}.$$

We now derive and solve equations for the quantities

$$Q_m(u) = \sum_k \Pr\{U_k \leqq u \quad \text{and} \quad N(U_k - 0) = m\},$$

which occur in the representation for the probability $\Pr\{N(t-0) = n\}$. Using the generating variable x and the abbreviation

$$P_y(x) = 1 + (x-1)e^{-\gamma y},$$

we find that

$$\sum_{n=0}^{N} x^n \Pr\{U_{k+1} \leqq t \quad \text{and} \quad N(U_{k+1} - 0) = n\} =$$

$$\sum_m \int_0^t \int_0^{t-u} P_y^{\,m+1-\delta_{mN}}(x)\, dA(y)\, d_u \Pr\{U_k \leqq u \quad \text{and} \quad N(U_k - 0) = m\}.$$

The series formed by adding all these equations up on the index k are absolutely convergent; hence no additional generating functions are needed, and we reach the equation:

$$\begin{aligned}
&\sum_{n=0}^{N} x^n \sum_{k=1}^{\infty} \Pr\{U_k \leqq t \quad \text{and} \quad N(U_k - 0) = n\} = \\
&\sum_{n=0}^{N} x^n \Pr\{u_1 \leqq t \quad \text{and} \quad N(u_1 - 0) = n\} \\
&+ \sum_m \int_0^t \int_0^{t-u} P_y^{\,m+1-\delta_{mN}}(x)\, dA(y)\, d_u \sum_k \Pr\{U_k \leqq u \quad \text{and} \\
&\qquad\qquad N(U_k - 0) = m\}.
\end{aligned} \tag{30}$$

This is an integral-functional equation for the function

$$\Psi(x,t) = \sum_n x^n \sum_k \Pr\{U_k \leqq t \quad \text{and} \quad N(U_k - 0) = n\},$$

which is closely related to the function $\psi(x,t,0)$ treated in Appendix A. In fact, when Ψ is absolutely continuous in t, then ψ is its derivative, and (22) is similar to (30) in the special case where the density ψ exists.

Equation (30) may be solved by the same method as (22), except that Laplace-Stieltjes transforms replace the ordinary Laplace integrals used for (22). We introduce the following notations:

$$Q_n(t) \quad \text{for} \quad \sum_k \Pr\{U_k \leqq t \quad \text{and} \quad N(U_k - 0) = n\},$$

$$B_n(t) \quad \text{for} \quad \sum_{j=n}^{N} \binom{j}{n} Q_j(t),$$

$$K_n(t) \quad \text{for} \quad \sum_{j=n}^{N} \binom{j}{n} \Pr\{u_1 \leqq t \quad \text{and} \quad N(u_1 - 0) = j\}.$$

When each of Q_n, B_n and K_n is absolutely continuous, the corresponding (lower case) quantities $q_n(t,0)$, $b_n(t)$ and $k_n(t)$ are the respective derivatives. Let the respective Laplace-Stieltjes transforms of Q_n, B_n and K_n be Q_n^*, B_n^* and K_n^*. Then (30) leads to the recurrence

$$B_n^* = a_n(s)\left\{B_n^* + B_{n-1}^* - \binom{N}{n-1} B_N^*\right\} + K_n^*. \tag{31}$$

The rest of the solution is in complete analogy with the solution for the b_n^*, q_n^* in Appendix A. We find

$$B_N(t) = Q_N(t),$$

$$B_0(t) = \sum_0^N Q_n(t) = \Psi(1,t).$$

The function $\Psi(1,t)$ satisfies the renewal equation

$$\Psi(1,t) = \int_0^t \Psi(1,t - y)\, dA(y) + G(t),$$

where $G = \text{distr}\,\{u_1\}$. The Laplace-Stieltjes transform of $\Psi(1,t)$ is

$$B_0^* = \int_0^\infty e^{-st}\, d\Psi(1,t) = \frac{\int_0^\infty e^{-st}\, dG(t)}{1 - A^*(s)} = \frac{K_0^*}{1 - A^*(s)}.$$

The solution of the recurrence (31) is

$$B_n^* =$$

$$\prod_0^n \frac{a_n(s)}{1 - a_n(s)} \left\{ B_0^* - \sum_{j=1}^n \left[\binom{N}{j-1} B_N^* - \frac{K_j^*}{a_j(s)} \right] \prod_0^{j-1} \frac{1 - a_i(s)}{a_i(s)} \right\},$$

where the first term of the products is taken to be one. The Q_n^* are given in terms of the B_n^* by the equation

$$Q_n^* = \sum_{j=0}^{N-n} (-1)^j \binom{n+j}{n} B_{n+j}^*.$$

In particular, the analog of (27) is

$$\begin{aligned} B_N^* = Q_N^* &= \int_0^\infty e^{-st}\, dt \sum_k \Pr\{U_k \leqq t \quad \text{and} \quad N(U_k - 0) = N\} \\ &= [1 - A^*(s)]^{-1} \qquad\qquad (32) \end{aligned}$$

$$\cdot \frac{K_0^* + K_1^* \dfrac{[1 - a_0(s)]}{a_1(s)} + \cdots + K_N^* \dfrac{[1 - a_0(s)] \cdots [1 - a_{N-1}(s)]}{a_1(s) \cdots a_N(s)}}{1 + \dbinom{N}{1} \dfrac{1 - a_1(s)}{a_1(s)} + \cdots + \dbinom{N}{N} \dfrac{[1 - a_1(s) \cdots [1 - a_N(s)]}{a_1(s) \cdots a_N(s)}}.$$

From the representation of $\Pr\{N(t - 0) = n\}$ it can be seen that the generating function of $N(t - 0)$ is

$$\begin{aligned} E\{x^{N(t-0)}\} &= E\{P_t^{N(0+)}(x) U(u_1 - t)\} \\ &\quad + \sum_n \int_0^t P_{t-u}^{1+n-\delta_{nN}}(x)[1 - A(t - u)]\, dQ_n(u), \end{aligned}$$

with $P_t(x) = 1 + (x - 1)e^{-\gamma t}$, and U the unit step at zero. It follows that the Laplace transform (with respect to t) of the generating function of $N(t - 0)$ is

$$\begin{aligned} \int_0^\infty e^{-st} E\{x^{N(t-0)}\}\, dt &= \int_0^\infty e^{-st} E\{P_t^{N(0+)}(x) U(u_1 - t)\}\, dt \\ &\quad + \sum_n Q_n^*(s) \int_0^\infty e^{-sy} P_y^{n+1-\delta_{nN}}(x)[1 - A(y)]\, dy. \end{aligned}$$

When $\lim E\{x^{N(t-0)}\}$ exists as $t \to \infty$, we can use Abel's theorem for the Laplace transform to evaluate the limits lim Pr $\{N(t - 0) = n\}$ explicitly. As $s \to 0$,

$$\lim sQ_n^*(s) = \lim \frac{sG_n^*(s)}{1 - F_{n,n}^*(s)}$$

$$= \frac{1}{E\{x_{n,n}\}}.$$

But from (32) we find

$$\lim sQ_n(s) = \frac{p_n}{\mu_1}$$

$$= \frac{p_n}{\int_0^\infty x\,dA(x)},$$

where p_n is the equilibrium probability that an arriving customer finds n trunks busy. (These probabilities have been studied in Takács[7] and Beneš,[11] *inter alia.*) Hence

$$E\{x_{n,n}\} = \text{mean recurrence time of } R_n$$

$$= \frac{\int_0^\infty x\,dA(x)}{p_n} = \frac{\mu_1}{p_n},$$

and, from Theorem 3,

$$\lim_{t\to\infty} \Pr\{N(t - 0) = n\} = \sum_{n-1 \leqq m \leqq N} p_m \int_0^\infty p_{m+1-\delta_{m,N},n}(u)\,\frac{[1 - A(u)]}{m_1}\,du$$

$$\lim_{t\to\infty} E\{x^{N(t-0)}\} = \sum_n p_n \int_0^\infty P_y^{\,n+1-\delta_{N,n}}(x)\,\frac{[1 - A(y)]}{m_1}\,dy,$$

with

$$P_y(x) = 1 + (x - 1)e^{-\gamma y}.$$

REFERENCES

1. Arrow, K. J., Karlin, S. and Scarf, H., *Studies in the Mathematical Theory of Inventory and Production*, Stanford Univ. Press, Palo Alto, Calif., 1958.
2. Palm, C., Intensitätsschwankungen im Fernsprechverkehr, Ericsson Tech., **44,** 1943.
3. Pollaczek, F., Lösung eines Geometrischen Wahrscheinlichkeitsproblems, Math. Zeits., **35,** 1932, p. 230.
4. Pollaczek, F., Généralization de la théorie probabiliste des systèmes téléphoniques sans dispositif d'attente, Compt. Rend., **236,** 1953, p. 1469.

5. Pollaczek, F., Détermination de différentes fonctions de répartition relatives à un groupe de lignes téléphoniques sans dispositif d'attente, Compt. Rend., **247,** 1958, p. 1826.
6. Pollaczek, F., Fonctions de répartition relatives à un groupe de lignes téléphoniques sans dispositif d'attente, Compt. Rend., **248,** 1959, p. 353.
7. Takács, L., On the Generalization of Erlang's Formula, Acta Math. Acad. Sci. Hung., **7,** 1956, p. 419.
8. Takács, L., On a Coincidence Problem Concerning Telephone Traffic, Acta Math. Acad. Sci. Hung., **9,** 1958, p. 45.
9. Takács, L., A Telefon-forgalom Elméletének Néhány Valószinüségszámitási Kerdeserol, Acta Magyar Tud. Akad. (Math. and Phys.), **8,** 1958, p. 151.
10. Cohen, J. W., The Full Availability Group of Trunks with an Arbitrary Distribution of the Interarrival Times and a Negative Exponential Holding-Time Distribution, Phillips Rep., 1956, p. 1.
11. Beneš, V. E., On Trunks with Negative Exponential Holding Times Serving a Renewal Process, B.S.T.J., **38,** 1959, p. 211.
12. Beneš, V. E., On the Covariance Function of A Simple Trunk Group, with Applications to Traffic Measurement, to be published.
13. Smith, W. L., Regenerative Stochastic Processes, Proc. Roy. Soc. (London), **232A,** 1955, p. 6.
14. Beneš, V. E., On Queues with Poisson Arrivals, Ann. Math. Stat., **28,** 1957, p. 670.
15. Wilkinson, R. I., Theories for Toll Traffic Engineering in the U. S. A., B.S.T.J., **35,** 1956, p. 421.
16. Szegö, G., *Orthogonal Polynomials*, American Mathematical Society, New York, 1938, p. 33.
17. Kosten, L., Over de invloed van herhaalde opreopen in de theorie der blokkeringskansen, De Ingenieur, **47** (Electrotechnik 14), 1947, p. 123.
18. Brockmeyer, E., The Simple Overflow Problem in the Theory of Telephone Traffic, Teleteknik, **5,** 1954, p. 361.
19. Kosten, L., Über Sperrungwahrscheinlichkeiten bei Staffelschaltungen, Elek. Nachrichten-tech., **14,** 1937, p. 5.
20. Cox, D. R., The Analysis of Non-Markovian Stochastic Processes by the Inclusion of Supplementary Variables, Proc. Camb. Phil. Soc., **51,** 1955, p. 433.
21. Smith, W. L., Asymptotic Renewal Theorems, Proc. Roy. Soc. (Edinburgh) **64A,** 1954, p. 9.

An Alternative Approach to the Realization of Network Transfer Functions: The *N*-Path Filter

By L. E. FRANKS and I. W. SANDBERG

(Manuscript received April 14, 1960)

A particular time-varying network consisting of several parallel transmission paths, each containing input and output modulators, is described and analyzed. It is shown that, under certain conditions, the network may be characterized by a transfer function. A particular form of this transfer function yields periodic filtering characteristics over a limited frequency band without employing distributed elements. Techniques are also presented for realizing highly selective band-pass filters without the use of magnetic elements. Some practical applications are discussed in detail and experimental verification is presented.

I. INTRODUCTION

The application of conventional design techniques to network problems in systems operating at relatively low frequencies often leads to impractical circuits. In addition, designs based on active *RC* techniques are frequently very sensitive to small changes in element values. Alternatively, a time-varying network approach to the solution of a wide class of such problems appears to be particularly promising.

The time-varying network described and analyzed in this paper consists of a parallel combination of *N* identical linear time-invariant networks, each operating between input and output modulators. Attention is focused upon several properties of this configuration that are of theoretical as well as practical importance. In particular, these properties include:

i. Periodic filtering characteristics can be obtained over a limited frequency band without employing distributed elements. The practical uses for this property include the realization of low-frequency comb filters.

ii. Narrow-band band-pass and band-elimination filters can be realized at very low frequencies by networks free from magnetic elements. The center frequency of these filters is electronically controllable.

iii. An exact low-pass to band-pass translated version of the constituent network transfer function can be realized. The low-pass to band-pass transformation technique can also be applied to driving-point immittances.

The network under consideration is shown in block diagram form in Fig. 1. The time functions $u(t)$, $v(t)$, $x_n(t)$ and $y_n(t)$ may be interpreted to be either voltages or currents. The input modulators (multipliers) operate on the input $u(t)$ to produce the inputs

$$x_n(t) = u(t)p[t - (n - 1)\tau]$$

to the N identical linear time-invariant networks with impulse response $h(t)$. The outputs $y_n(t)$ are passed through output modulators to form path outputs $v_n(t)$. The final output $v(t)$ is the sum of the path outputs. The time functions $p[t - (n - 1)\tau]$ and $q[t - (n - 1)\tau]$ are periodic with period T, where $T = N\tau$.

In the next section the general input-output relationship for the N-path configuration is developed and discussed. The following sections are concerned with features associated with particular types of modulating functions. Some practical applications are discussed in detail and experimental verification is presented.

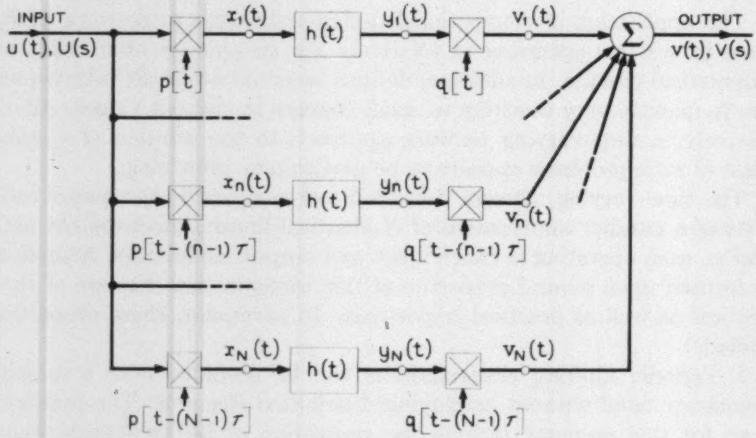


Fig. 1. — The N-path configuration.

II. GENERAL PROPERTIES OF THE N-PATH CONFIGURATION

2.1 *General Input-Output Relationship*

In this section we derive the relationship between $V(s)$ and $U(s)$, the network's frequency domain output and input.†

The periodic functions $p(t)$ and $q(t)$ can be expressed by their complex Fourier series:

$$\begin{aligned} p(t) &= \sum_{m=-\infty}^{m=+\infty} P_m e^{j\omega_0 m t}, \\ q(t) &= \sum_{l=-\infty}^{l=+\infty} Q_l e^{j\omega_0 l t}, \end{aligned} \tag{1}$$

where $\omega_0 = 2\pi/T = 2\pi/N\tau$. It is convenient to define

$$\begin{aligned} p_n(t) &= p[t - (n-1)\tau], \\ q_n(t) &= q[t - (n-1)\tau]. \end{aligned} \tag{2}$$

Since multiplication in the time domain corresponds to convolution in the frequency domain, it follows that

$$V(s) = \sum_{n=1}^{N} V_n(s) = \sum_{n=1}^{N} Y_n(s) \otimes Q_n(s). \tag{3}$$

Using the relation

$$J(s) \otimes \frac{1}{s - \alpha} = J(s - \alpha) \tag{4}$$

and (1), (2) and (3), we obtain

$$V(s) = \sum_{n=1}^{N} \sum_{l=-\infty}^{l=+\infty} Q_l e^{-j\omega_0(n-1)l\tau} X_n(s - jl\omega_0) H(s - jl\omega_0), \tag{5}$$

where

$$X_n(s)H(s) = Y_n(s). \tag{6}$$

Similarly,

$$X_n(s) = U(s) \otimes P_n(s) \tag{7}$$

and

$$X_n(s - jl\omega_0) = \sum_{m=-\infty}^{m=+\infty} P_m e^{-j\omega_0(n-1)m\tau} U[s - j(m + l)\omega_0].$$

---

† The time function and its Laplace transform are denoted, in accordance with the usual notation, by lower and upper case letters respectively.

Substituting (7) into (5) gives

$$V(s) = \sum_{n,l,m} Q_l P_m e^{-j\omega_0(n-1)(l+m)\tau} H(s - jl\omega_0) U[s - j(m + l)\omega_0]. \tag{8}$$

The summation over n is the following geometric series:

$$\sum_{n=1}^{N} e^{-j\omega_0(n-1)(l+m)\tau} \begin{cases} = N, & l + m = kN, \\ = 0, & \text{otherwise,} \end{cases} \tag{9}$$

where k is an integer. Using (9), we obtain

$$V(s) = N \sum_{k,l} Q_l P_{kN-l} H(s - jl\omega_0) U(s - jkN\omega_0). \tag{10}$$

It is convenient to write (10) in the form

$$V(s) = \sum_{k=-\infty}^{k=+\infty} F(k,s) U(s - jkN\omega_0), \tag{11}$$

$$F(k,s) = N \sum_{l=-\infty}^{l=+\infty} Q_l P_{kN-l} H(s - jl\omega_0). \tag{12}$$

Expressions (11) and (12) constitute the general input-output relationship for the N-path structure.

2.2 *Transfer Function for N-Path Configuration*

The quantity $F(k,s)$ in (11) and (12) completely characterizes the time-varying network of Fig. 1. It describes operationally the relation between the input signal and output signal, as is shown symbolically in Fig. 2(a). In this sense, $F(k,s)$ may be considered analogous to the characterization of a constant-parameter network in terms of a transfer function. A feature of the N-path configuration of particular interest from the network theory viewpoint is that, with certain band-limiting restrictions on the input and output signals, a transfer function relation between input and output can be derived. It is this property that will be investigated in the remainder of the paper.

If $U(s)$ evaluated on the $j\omega$-axis essentially vanishes outside the interval $|\omega| < N\omega_0/2$, it follows that

$$V(j\omega) = F(0,j\omega) U(j\omega) \qquad \text{in } |\omega| < \frac{N\omega_0}{2}. \tag{13}$$

Furthermore, if $V(j\omega)$ vanishes outside the interval $|\omega| < N\omega_0/2$, then $V(s)$ and $U(s)$ can be related by a transfer function $T(s)$:

$$T(s) = \frac{V(s)}{U(s)},$$

where

$$\begin{aligned} T(j\omega) &= F(0,j\omega) \qquad \text{in } |\omega| < \frac{N\omega_0}{2}, \\ &= 0 \qquad\qquad\quad \text{in } |\omega| > \frac{N\omega_0}{2}. \end{aligned} \tag{14}$$

These band-limiting constraints can be accomplished by preceding and following the time-varying network with ideal low-pass filters having cutoff at $\omega_c = N\omega_0/2$. With the addition of these low-pass filters, the time-varying network is equivalent to a constant-parameter network having a transfer function, $F(0,s)$, preceded and followed by ideal low-pass filters, as shown in Fig. 2(b).

An alternate expression for the transfer function will be developed in the following equations. This expression leads to a closed form for $F(0,s)$.

From (12),

$$F(0,s) = N \sum_{l=-\infty}^{\infty} P_{-l}Q_l H(s - jl\omega_0). \tag{15}$$

This can be written as the Laplace transform of the product of the im-

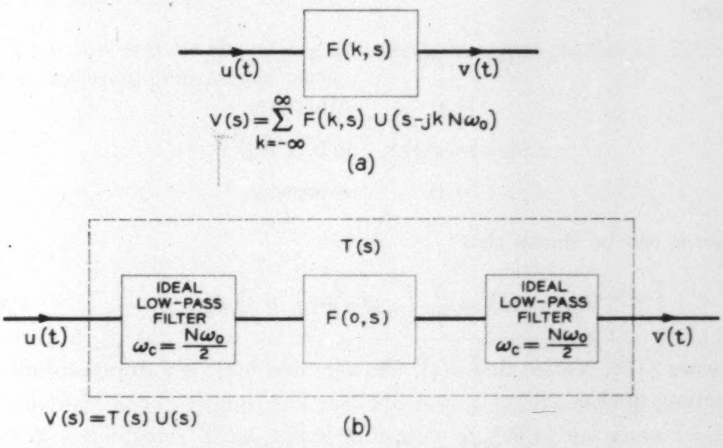

F(k, s)
u(t)
v(t)
$V(s)=\sum_{k=-\infty}^{\infty} F(k,s)\, U(s-jkN\omega_0)$
(a)
T(s)
IDEAL LOW-PASS FILTER $\omega_c=\frac{N\omega_0}{2}$
F(o, s)
IDEAL LOW-PASS FILTER $\omega_c=\frac{N\omega_0}{2}$
u(t)
v(t)
V(s) = T(s) U(s)
(b)

Fig. 2 — (a) Symbolic representation of $F(k,s)$; (b) equivalent constant-parameter network.

pulse response of the component networks, $h(t)$, and a periodic function with period T:

$$\begin{aligned} F(0,s) &= \mathfrak{L}\left[h(t)N\sum_{l=-\infty}^{\infty} P_{-l}Q_l e^{+jl\omega_0 t}\right] \\ &= \mathfrak{L}\left[h(t)\sum_{k=-\infty}^{\infty} r(t-kT)\right], \end{aligned} \tag{16}$$

where

$$\sum_{k=-\infty}^{\infty} r(t-kT) = N\sum_{l=-\infty}^{\infty} P_{-l}Q_l e^{+jl\omega_0 t}. \tag{17}$$

The pulse $r(t)$ depends only on the modulating functions and not on the response characteristics of the component networks. The identification of $r(t)$ with $p(t)$ and $q(t)$ is not unique. However, a particularly useful relation is obtained by considering $p(t)$ and $q(t)$ to be represented by infinite pulse trains wherein each pulse assumes the shape of one period of the modulating functions; that is,

$$\begin{aligned} p(t) &= \sum_{k=-\infty}^{\infty} a(t-kT), \\ q(t) &= \sum_{k=-\infty}^{\infty} b(t-kT), \end{aligned} \tag{18}$$

where

$$\begin{aligned} a(t) &= p(t) && \text{in } 0 \leqq t \leqq T, \\ &= 0 && \text{otherwise;} \\ b(t) &= q(t) && \text{in } 0 \leqq t \leqq T, \\ &= 0 && \text{otherwise.} \end{aligned} \tag{19}$$

Then it can be shown that

$$r(t) = \frac{N}{T}\int_0^T a(y)b(y+t)\,dy \tag{20}$$

satisfies (17). Notice that $r(t)$, like $a(t)$ and $b(t)$, is a duration-limited function, in that

$$r(t) = 0 \qquad \text{for } |\,t\,| \geqq T. \tag{21}$$

Since the Laplace transform of a product of time functions is given by the convolution of their transforms, (16) becomes

$$F(0,s) = H(s) \otimes \mathfrak{L}\left[\sum_{k=-\infty}^{\infty} r(t - kT)\right] \tag{22}$$

and

$$\mathfrak{L}\left[\sum_{k=-\infty}^{\infty} r(t - kT)\right] = R(s) + \frac{N}{T}\sum_{k=1}^{\infty} A(-s)B(s)e^{-skT}, \tag{23}$$

where

$$\begin{aligned} R(s) &= \int_0^T r(t)e^{-st}\,dt, \\ A(s) &= \int_0^T a(t)e^{-st}\,dt, \\ B(s) &= \int_0^T b(t)e^{-st}\,dt. \end{aligned} \tag{24}$$

The terms in k form a geometric series that is readily summed, so that

$$F(0,s) = H(s) \otimes \left[R(s) + \frac{\frac{N}{T}A(-s)B(s)e^{-sT}}{1 - e^{-sT}}\right]. \tag{25}$$

2.3 *Transfer Function for Rational $H(s)$*

If we now assume that $H(s)$ is rational in s and regular at infinity, then, assuming only simple poles,

$$H(s) = c_0 + \sum_{i=1}^{M} \frac{c_i}{s - s_i}. \tag{26}$$

From (25),

$$F(0,s) = $$

$$c_0r(0) + \sum_{i=1}^{M} c_i\left[R(s - s_i) + \frac{\frac{N}{T}A(s_i - s)B(s - s_i)e^{-(s-s_i)T}}{1 - e^{-(s-s_i)T}}\right]. \tag{27}$$

The functions $R(s)$ and $A(-s)B(s)$ have no singularities in the finite part of the s-plane. Thus, the singularities of $F(0,s)$ are given by the zeros of $1 - e^{-(s-s_i)T}$, which lie equally spaced at intervals of $2\pi/T$ on lines parallel to the $j = \omega$ axis.

III. SPECIFIC TYPES OF MODULATING FUNCTIONS

In this section the properties of the N-path configuration are examined for specific types of modulation that reveal particularly interesting properties of the structure.

3.1 *Sinusoidal Modulation*

Suppose that the modulating functions possess only a finite number of sinusoidal components:

$$\begin{aligned} p(t) &= \sum_{m=-M}^{M} P_m e^{j\omega_0 mt}, \\ q(t) &= \sum_{m=-M}^{M} Q_m e^{j\omega_0 mt}, \end{aligned} \tag{28}$$

where

$$P_{-m} = P_m^* \qquad \text{and} \qquad Q_{-m} = Q_m^*.$$

The case for $N > 2M$ deserves special attention, for then

$$Q_l P_{kN-l} = 0 \qquad \text{for } k \neq 0,$$

and, from (11) and (12),

$$\frac{V(s)}{U(s)} = F(0,s). \tag{29}$$

Therefore, the network exhibits a transfer function $T(s)$ for $N > 2M$, without band-limiting restrictions, which is given by

$$T(s) = F(0,s).$$

Note that the transfer function is a finite sum of frequency-translated versions of $H(s)$. In particular, when $P_m, Q_m = 0$ for $|m| \neq 1$, we have

$$T(s) = N[a_1 H(s - j\omega_0) + a_1^* H(s + j\omega_0)], \tag{30}$$

where

$$a_1 = Q_1 P_{-1},$$

a "low-pass to band-pass transformation" of the transfer function $H(s)$†.

---

† This result can also be obtained with only two parallel paths.[1] Single sinusoid modulating functions are employed, the two functions in one path being in phase with each other and in quadrature with the functions in the other path. A similar configuration has been considered by Hines and Desoer in unpublished work.

A particularly difficult practical network problem is the low-frequency realization of highly selective band-pass filters. Procedures that avoid the use of magnetic elements are inviting, but active *RC* techniques often lead to a high degree of transfer function sensitivity to both the active and passive parameters. An alternate approach based on (30) appears to be attractive, and should provide a considerable increase in the degree of immunity from network parameter variations. Implementation of a similar approach is discussed in more detail in Section 5.2.

The transfer function poles of a passive *RC* network are distinct and on the negative-real axis of the complex-frequency plane. Consequently, if $H(s)$ is the transfer function of an *RC* network, the over-all transfer function $T(s)$ of (30) can have only distinct pairs of complex-conjugate poles with identical imaginary parts. It is desirable to circumvent this restriction without employing magnetic or active elements. It is sufficient to consider the synthesis of the transfer function

$$T(s) = \frac{N(s)}{D(s)}, \tag{31}$$

where $T(s)$ has only simple complex-conjugate poles, since transfer functions with multiple-order poles can be realized as the product of transfer functions having only simple poles. We assume that $T(s)$ is stable and regular at infinity. Equation (31) can be expressed as

$$T(s) = K_\infty + \sum_{i=1}^{M} \frac{b_i}{s + \sigma_i - j\omega_i} + \frac{b_i^*}{s + \sigma_i + j\omega_i}. \tag{32}$$

From (30), each of the series terms can be separately realized with the passive transfer function $H_i(s) = 1/(s + \sigma_i)$. Evidently we require

$$b_i = NQ_{1i}P_{-1i}. \tag{33}$$

Hence, a realization of (32) consists of M similar sections in parallel, with an additional section that realizes the constant term. The main objection to this realization technique is that a large number of modulators may be required, but it demonstrates that any transfer function that is regular at infinity and stable can be realized with sinusoidal modulators, a source of modulating frequencies and simple passive *RC* structures.

While this paper is primarily concerned with the synthesis of transfer functions, it is worthwhile to sacrifice some degree of continuity here to point out the relevance and extension of the preceding discussion to the synthesis of driving-point impedances. The results of this section apply also to the case where $U(s)$ and $V(s)$ are interpreted to correspond to the transforms of voltage and current at the same port. The forms taken by the network for this special application are shown in Fig. 3. Suppose, for example, that the nth two-port network in Fig. 3(a) is characterized by

$$\begin{aligned} i_n'(t) &= p[t - (n - 1)\tau]e_n(t), \\ i_n(t) &= q[t - (n - 1)\tau]e_n'(t), \end{aligned} \tag{34}$$

where $p(t)$ and $q(t)$ are given by (28). The driving-point admittance

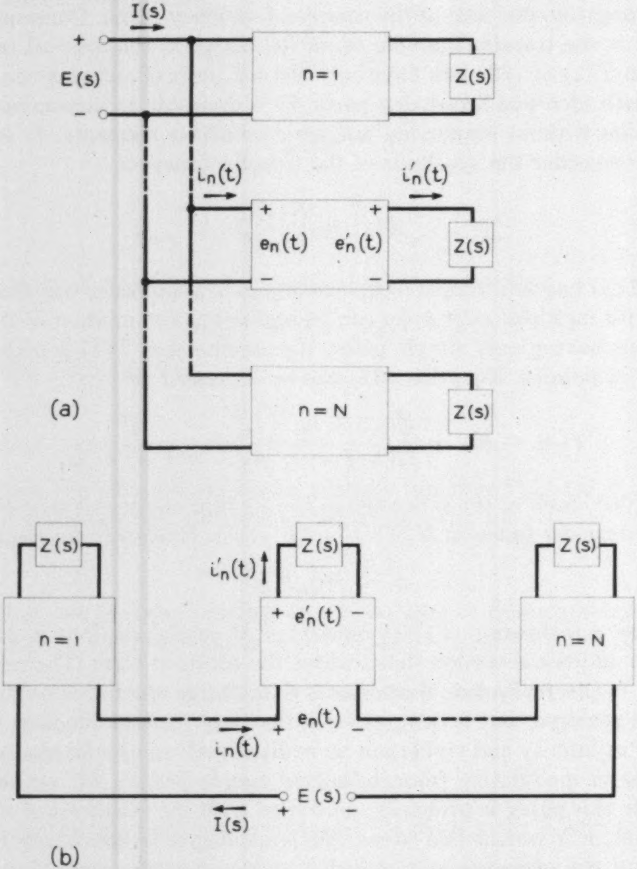

I(s)
+
E(s)
−
n=1
Z(s)
$i_n(t)$
$i_n'(t)$
+
+
$e_n(t)$
$e_n'(t)$
−
−
Z(s)
(a)
n=N
Z(s)
Z(s)
Z(s)
Z(s)
$i_n'(t)$
n=1
+ $e_n'(t)$ −
n=N
$e_n(t)$
+
−
$i_n(t)$
E(s)
+
−
I(s)
(b)

Fig. 3 — Forms taken by network when $U(s)$ and $V(s)$ are interpreted as corresponding to transforms of voltage and current at same port.

$Y_{in} = I(s)/E(s)$ is given by $F(0,s)$, with $H(s)$ replaced by $Z(s)$. That is,

$$Y_{in}(s) = NP_0Q_0Z(s) + N\sum_{m=1}^{M}[a_mZ(s - jm\omega_0) + a_m^*Z(s + jm\omega_0)], \tag{35}$$

where

$$a_m = Q_mP_{-m} \qquad \text{and} \qquad N > 2M.$$

For the special case where P_m, $Q_m = 0$ for $|m| \neq 1$ and a_1 is real,

$$Y_{in}(s) = Na_1[Z(s - j\omega_0) + Z(s + j\omega_0)]. \tag{36}$$

For example, if $Z(s) = 1/sC$,

$$Y_{in}(s) = \frac{2Na_1}{C}\frac{s}{s^2 + \omega_0^2}, \tag{37}$$

the admittance of an inductor and capacitor in series.

As in the transfer function case, (36) (and the analogous relations for the following three other networks discussed here) can be realized with only two parallel paths. Single sinusoid modulating functions are employed, the two functions in each two-port network being in quadrature with the corresponding functions in the other two-port network.

If the two-port networks in Fig. 3(a) are characterized by

$$\begin{aligned} e_n'(t) &= p[t - (n - 1)\tau]e_n(t), \\ i_n(t) &= q[t - (n - 1)\tau]i_n'(t), \end{aligned} \tag{38}$$

we obtain

$$Y_{in}(s) = NP_0Q_0Y(s) + N\sum_{m=1}^{M}[a_mY(s - jm\omega_0) + a_m^*Y(s + jm\omega_0)], \tag{39}$$

where

$$Y(s) = \frac{1}{Z(s)}.$$

The two dual networks take the form shown in Fig. 3(b).

3.2 *Jump Modulation*

The physical implementation of the transfer function of (15) can be accomplished without the difficulties normally encountered in the realiza-

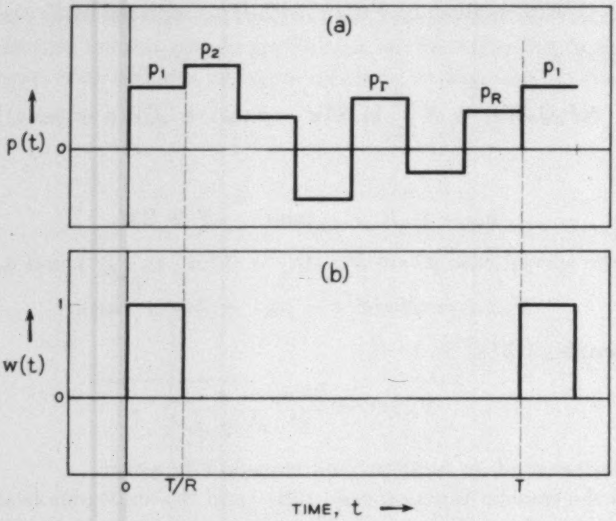


Fig. 4 — (a) Modulating function $p(t)$ with R jumps in fundamental interval T; (b) periodic switching function $w(t)$.

tion of accurate multiplier circuitry by means of a scheme called *jump modulation*. This scheme uses modulating functions having a finite number of equally spaced discontinuities or jumps in each fundamental interval. The functions assume a constant value between jumps. Modulators of this type can be realized by conventional switching techniques. Suppose that the modulating function $p(t)$ has R jumps in the fundamental interval T, as shown in Fig. 4(a):

$$p(t) = \sum_{r=1}^{R} p_r w\left[t - (r-1)\frac{T}{R}\right] = \sum_{m=-\infty}^{\infty} P_m e^{j(m2\pi/T)t} \tag{40}$$

where $w(t)$ is the periodic switching function shown in Fig. 4(b).

The Fourier coefficients of $p(t)$ are given by

$$P_m = \frac{1}{T}\int_0^T p(t)\, e^{-j(m2\pi/T)t}\, dt = \frac{1}{T}\sum_{r=1}^{R} p_r \int_{(r-1)(T/R)}^{r(T/R)} e^{-j(m2\pi/T)}\, dt. \tag{41}$$

Thus, the sequence of values P_m is given by a linear transformation of the sequence of values, p_r:

$$P_m = \sum_{r=1}^{R} \frac{e^{j(m\pi/R)} \sin \dfrac{m\pi}{R}}{m\pi}\, e^{-j(2\pi/R)mr} p_r\,. \tag{42}$$

For design purposes, the inverse of this transformation is desired in order that an appropriate set, p_r, can be determined from an arbitrarily prescribed set, P_m. Obviously, only R values can be independently prescribed for the complex numbers P_m and, since $p(t)$ is real, it is always required that $P_{-m} = P_m^*$ and P_0 be real. Hence, for example, a set of values p_r can be found such that all the Fourier coefficients P_m can be arbitrarily specified for $|m| \leqq R/2$. In this case, the inverse transformation corresponding to (42) is relatively simple:[†]

$$p_r\ (R \text{ odd}) = \sum_{m=-(R-1)/2}^{(R-1)/2} \frac{e^{-j(m\pi/R)}\,\dfrac{m\pi}{R}}{\sin\dfrac{m\pi}{R}}\, e^{j(2\pi/R)rm} P_m . \tag{43}$$

When R is even,[‡]

$$p_r\ (R \text{ even}) = \sum_{m=-R/2+1}^{R/2-1} \frac{e^{-j(m\pi/R)}\,\dfrac{m\pi}{R}}{\sin\dfrac{m\pi}{R}}\, e^{j(2\pi/R)rm} P_m - j\,\frac{\pi}{2}(-1)^r P_{R/2} . \tag{44}$$

A case of particular interest in the N-path configuration is for $R = N$, when the jumps occur simultaneously in all paths and a common timing source can be used for operating the switches. If the bandwidth of the component networks is sufficiently small compared to ω_0, the transfer function can be expressed approximately in terms of the first $N/2$ Fourier coefficients:

$$T(j\omega) \cong N \sum_{m=-N/2}^{N/2} Q_m P_{-m} H(j\omega - jm\omega_0), \tag{45}$$

where all the values of either P_m or Q_m or both can be arbitrarily chosen.

3.3 *Pulse Modulation*

A special case of jump modulation of considerable practical importance is for the set $\{p_1 = 1,\ p_2 = p_3 = \cdots = p_{R_1} = 0\}$ and $\{q_1 = 1,\ q_2 = q_3 = \cdots = q_{R_2} = 0\}$, so that

---

[†] See Appendix A for derivation of the inverse transformation.

[‡] From (42), it is seen that the real part of $P_{R/2}$ must vanish, since

$$P_{R/2} = e^{j(\pi/2)} \frac{\sin(\pi/2)}{R(\pi/2)} \sum_{r=1}^{R} (-1)^r p_r .$$

$$\begin{aligned} p(t) &= w_1(t), \\ q(t) &= w_2(t), \end{aligned} \tag{46}$$

If the input generator, $u(t)$, is a current source, each modulator in Fig. 1 can be replaced by a simple switch. In fact, when R_1 and $R_2 > N$, the entire set of input and output modulators would then be equivalent to a pair of N-contact rotary switches on a common shaft rotating at a rate of $1/T$ cps. The dwell time at each contact of the input and output switches is given by $d_1 = T/R_1$ and $d_2 = T/R_2$, respectively. In this case, the switches are essentially signal-sampling devices, hence the general configuration using this type of modulation will hereafter be referred to as the *N-path sampled-data network.*

Besides being relatively simple to implement, the N-path sampled-data network has some very interesting transfer function characteristics. If the component networks have a low-pass characteristic with bandwidth small compared to ω_0, the transfer function for large N will appear as a sequence of narrow, equally spaced passbands of identical shape and nearly equal height, centered at integral multiples of ω_0. This corresponds to the so-called "comb filter" characteristic, which is frequently employed in the detection of periodic signals immersed in wide band noise. Furthermore, it will be shown that the function $F(0,s)$ becomes periodic on the $j\omega$-axis as the dwell times d_1 and d_2 approach zero. When $H(s)$ is rational in s, this periodic function is of the form generally associated with the network functions of circuits containing resistors and ideal delay lines.

IV. TRANSFER FUNCTION FOR N-PATH SAMPLED-DATA NETWORK

The expression for $F(0,s)$ in terms of $r(t)$ as given in (22) is especially convenient for finding the transfer function of the N-path sampled-data network. Also, if $H(s)$ is rational in s and regular at infinity, then (27) gives an exact closed-form expression for $F(0,s)$.

Suppose, for example, that $d_1 = d_2 = d < T$. Then $r(t)$ is simply the triangular pulse,

$$\begin{aligned} r(t) &= \frac{N}{T}(d - |t|) \qquad \text{in } |t| \leqq d, \\ &= 0 \qquad\qquad\qquad\quad \text{otherwise,} \end{aligned} \tag{47}$$

so

$$R(s) = \frac{N}{T}\left[\frac{ds - (1 - e^{-sd})}{s^2}\right] \tag{48}$$

and

$$\begin{aligned} A(-s)B(s) &= \frac{(1 - e^{sd})}{-s}\frac{(1 - e^{-sd})}{s} \\ &= \frac{e^{sd} - 2 + e^{-sd}}{s^2}. \end{aligned} \tag{49}$$

Then, assuming $H(s)$ to be in the form of (26), the transfer function is obtained directly from (27):

$$F(0,s) = \frac{c_0Nd}{T} + \frac{N}{T}\sum_{i=1}^{M}\left(\frac{c_i}{\lambda_i^2}\right)\frac{(e^{-\lambda_i d} - 1 + \lambda_i d) + (e^{\lambda_i d} - 1 - \lambda_i d)e^{-\lambda_i T}}{1 - e^{-\lambda_i T}}, \tag{50}$$

where

$$\lambda_i = s - s_i .$$

When $| \lambda_i d | \ll 1$, the transfer function can be approximated by a function that is periodic for values of s on any line parallel to the $j\omega$-axis. If the first three terms in the power series expansion for $e^{\lambda_i d}$ and $e^{-\lambda_i d}$ are retained, then

$$F(0,s) \cong \frac{c_0Nd}{T} + \frac{Nd^2}{2T}\sum_{i=1}^{M} c_i \frac{1 + e^{-(s-s_i)T}}{1 - e^{-(s-s_i)T}} \tag{51}$$

for

$$| s - s_i | d \ll 1.$$

The relation (51) can be obtained in a different manner by application of conventional sampled-data techniques.[2] These techniques provide an alternate approach worthy of investigation, since they lead to a simple single-path sampled-data network, which is equivalent to the N-path sampled-data network. The approximation involved in this method of analysis consists of replacing sampling switches by impulse modulators (IM), as shown in Fig. 5(a). The train of narrow rectangular modulating pulses, $w(t)$, has been replaced by an impulse train, where the magnitude of each impulse is equal to the area of the corresponding rectangular pulse. Hence,

$$p(t) \cong d_1 \sum_{k=-\infty}^{\infty} \delta(t - kT), \tag{52}$$

$$q(t) \cong d_2 \sum_{k=-\infty}^{\infty} \delta(t - kT). \tag{53}$$

Then,

$$P_m = \frac{d_1}{T} \qquad \text{for all } m$$

and

$$Q_m = \frac{d_2}{T} \qquad \text{for all } m.$$

In this case, $F(k,s)$ in (12) is independent of k and

$$F(k,s) = G(s) = N \frac{d_1 d_2}{T^2} \sum_{l=-\infty}^{\infty} H(s - jl\omega_0). \tag{54}$$

Then,

$$V(s) = G(s) \left[\sum_{m=-\infty}^{\infty} U(s - jkN\omega_0) \right]. \tag{55}$$

This input-output relation is identical to that of a single-path sampled-data network having an input impulse modulator with sampling interval, $\tau = T/N$, followed by a network with a transfer function, $\tau G(s)$, which, when $s = j\omega$, is periodic with period ω_0. The periodic property of the

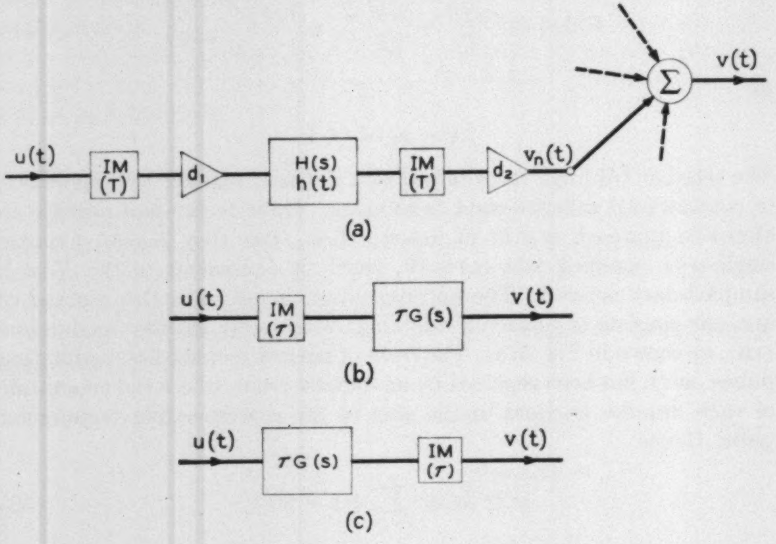


Fig. 5 — (a) Approximate representation of N-path sampled-data network; (b) and (c) equivalent single-path networks.

network following the impulse modulator in Fig. 5(b) allows the derivation of another equivalent network with the impulse modulator at the output, as shown in Fig. 5(c). These simple equivalent networks are very convenient for analysis purposes when one or more N-path configurations are component parts of a larger system.

The Fourier coefficients for the expression of $G(s)$ when $s = j\omega$ are obtained directly from the sample values, $h(rT)$, of the impulse response of one of the component networks. If

$$G(j\omega) = \sum_{i=-\infty}^{\infty} g_r \, e^{jr(2\pi\omega/\omega_0)},$$

then

$$\begin{aligned} g_r &= \frac{1}{\omega_0} \int_{-\omega_0/2}^{\omega_0/2} G(j\omega) \, e^{-j(r2\pi\omega/\omega_0)} \, d\omega \\ &= \left(\frac{Nd_1d_2}{T^2}\right) \frac{1}{\omega_0} \int_{-\omega_0/2}^{\omega_0/2} \sum_{l=-\infty}^{\infty} H(j\omega - jl\omega_0) \, e^{-j(r2\pi\omega/\omega_0)} \, d\omega \\ &= \left(\frac{Nd_1\,d_2}{T^2}\right) (T) \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} H(j\omega) \, e^{-jrT\omega} \, d\omega\right], \end{aligned} \tag{56}$$

$$g_r = \frac{Nd_1\,d_2}{T} \, h(-rT). \tag{57}$$

The integral in (56) is the Fourier inversion integral for $h(t)$. At discontinuities in $h(t)$ the inversion integral gives the mean value of the right- and left-hand limits at the discontinuity. Hence, for physically realizable component networks.

$$G(s) = \frac{Nd_1d_2}{T} \left[\frac{h(0+)}{2} + \sum_{r=1}^{\infty} h(rT) \, e^{-rsT}\right]. \tag{58}$$

This expansion is particularly useful when $H(s)$ is a rational function of s. In this case, the series can be summed and $G(s)$ is given in closed form. Assuming $H(s)$ has simple poles, then†

$$h(t) = \sum_{i=1}^{M} c_i \, e^{s_i t} \qquad \text{for} \quad t \geqq 0 \tag{59}$$

and, from (58),

$$G(s) = \frac{Nd_1d_2}{T} \sum_{i=1}^{M} c_i \left(\frac{1}{2} + \sum_{r=1}^{\infty} e^{r(s_i - s)T}\right). \tag{60}$$

---

† In this analysis we require that $H(s) \to 0$ as $s \to \infty$ ($c_0 = 0$) since the Laplace transform for a product of impulse functions is not defined.

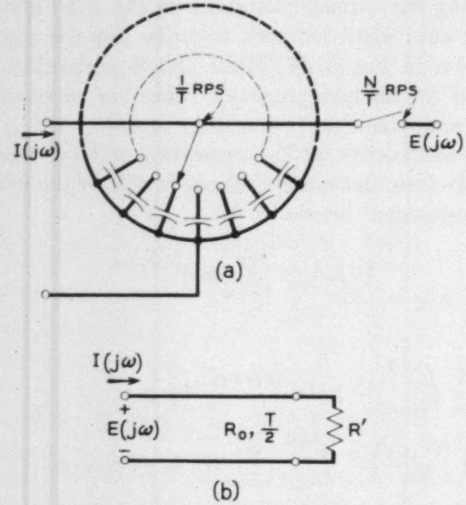


Fig. 6 — The *N*-capacitor element.

The sum over r in (60) is a geometric series and can be written in closed form, so that

$$G(s) = \frac{Nd_1d_2}{2T} \sum_{i=1}^{M} c_i \frac{1 + e^{(s_i-s)T}}{1 - e^{(s_i-s)T}}. \tag{61}$$

Note the equivalence between (61) obtained by conventional sampled-data techniques and the direct approximation of (51) to the transfer function of the N-path sampled-data network.

A simple example that illustrates the application of the preceding techniques is the case where each component network is a single capacitor, as shown in Fig. 6(a).† Capacitor loss is accounted for by the inclusion of a resistance, R_c, across each capacitor. The relation between input current and output voltage is represented by $G(s)$ in (61), where

$$H(s) = \frac{\dfrac{1}{C}}{s + \dfrac{1}{R_cC}}, \tag{62}$$

---

† This case has been described in the literature.[3,4]

so that

$$\begin{aligned} M &= 1, \\ s_1 &= -\frac{1}{R_eC}, \\ c_1 &= \frac{1}{C}, \\ d_1 &= d_2 = d \end{aligned} \tag{63}$$

and

$$G(s) = R_0\left[\frac{1 + \rho e^{-sT}}{1 - \rho e^{-sT}}\right], \tag{64}$$

where

$$\rho = e^{-T/R_eC}$$

and

$$R_0 = \frac{Nd^2}{2TC}.$$

The expression of (64) is equal to the driving-point impedance of a length of lossless transmission line of characteristic impedance, R_0, terminated at a distance corresponding to an electrical delay of $T/2$ seconds. The termination is characterized by a reflection coefficient $\rho = e^{-T/R_eC}$, or equivalently, by a resistance, R', where

$$R' = R_0\left(\frac{1 + \rho}{1 - \rho}\right). \tag{65}$$

If the capacitors are lossless, then

$$G(s) = \frac{Nd^2}{2TC}\left(\frac{1 + e^{-sT}}{1 - e^{-sT}}\right) = R_0 \coth\frac{sT}{2}, \tag{66}$$

which is equal to the driving-point impedance of the same length of lossless transmission line with open-circuit termination.

V. SOME PRACTICAL APPLICATIONS FOR THE *N*-PATH SAMPLED-DATA NETWORK

5.1 *Delay Network*

The transcendental nature of $G(s)$ of (66) for the N-capacitor element suggests the possibility of realizing an all-pass constant-delay characteristic over a limited bandwidth without the use of inductors. One of several possible configurations for accomplishing this is the simple feedback network shown in Fig. 7, where the N-capacitor element is contained in the feedback path.

For the analysis of this circuit, it is assumed that the forward gain, μ, is sufficiently large that the error voltage, $v_1 + v_3$, is essentially zero. Note that

$$v_2(t) = v_3(t) + R_2 i(t) \tag{67}$$

and, hence, because of the low-pass filter at the output, only the components of $V_3(j\omega)$ and $I(j\omega)$ in the frequency range $|\omega| \leqq N\omega_0/2$ are of interest. If $V_1(j\omega)$ [and hence $V_3(j\omega)$] is limited to this same bandwidth, then

$$I(s) = Y(s)V_3(s)$$

and

$$\frac{V_4}{V_1}(s) = -[1 + R_2 Y(s)] + K \tag{68}$$

over the frequency band of interest, and $Y(s)$ is a function of the N-path type.

The constant-delay characteristic is obtained by making the R_1C time constant very small compared to the contact dwell time, d. Roughly speaking, this means that the capacitors charge up to the applied voltage in the time interval d, during which their respective switches are closed, and the resulting current flow is a sequence of narrow exponentially decaying pulses occurring τ seconds apart. An approximate representation

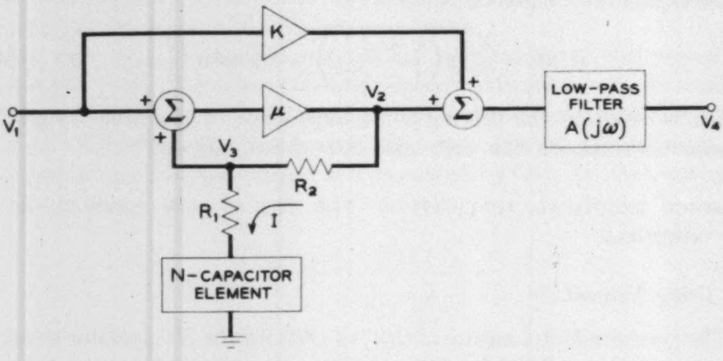


Fig. 7 — Constant-delay network.

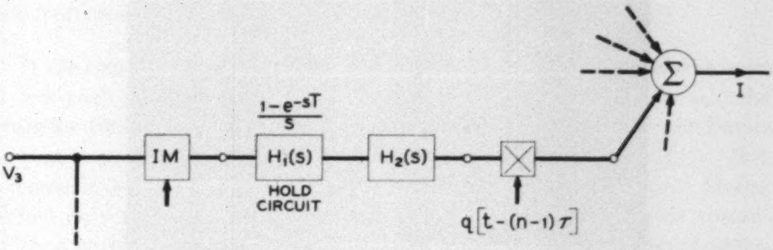


Fig. 8 — N-path network used for describing the behavior of the constant-delay network.

of this behavior in terms of the general N-path configuration is shown in Fig. 8. The applied voltage, V_3, is sampled with an impulse modulator at the time of the nth switch closure and held at this value for T seconds by means of the hold circuit, $H_1(s)$. The current flowing in the series combination of R_1 and C in response to the applied voltage steps is obtained by means of the transfer function

$$H_2(s) = \frac{1}{R_1}\frac{s}{s+\alpha}, \tag{69}$$

where

$$\alpha = \frac{1}{R_1C} \gg d.$$

The transfer function $Y(s)$ is obtained from (25), where

$$\begin{aligned} A(s) &= 1, \\ R(s) &= \frac{N}{T}B(s) = \frac{N}{T}\left(\frac{1-e^{-sd}}{s}\right). \end{aligned} \tag{70}$$

Then,

$$\begin{aligned} Y(s) &= \frac{N}{TR_1}\frac{1-e^{-sT}}{s+\alpha} \otimes \left[\frac{1-e^{-sd}}{s} + \frac{(1-e^{-sd})\,e^{-sT}}{s(1-e^{-sT})}\right] \\ &= \frac{N(1-e^{-sT})}{TR_1(s+\alpha)}\left\{1 - e^{-(s+\alpha)d} + \frac{[1-e^{-(s+\alpha)d}]\,e^{-(s+\alpha)T}}{1-e^{-(s+\alpha)T}}\right\}. \end{aligned} \tag{71}$$

Since $\alpha d \gg 1$, terms involving the factor $e^{-\alpha d}$ are neglected, and (71) is approximated by

$$Y(s) \cong \frac{N}{TR_1\alpha}(1-e^{-sT}) \tag{72}$$

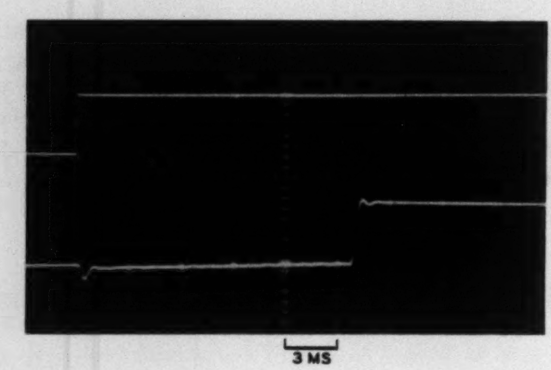


Fig. 9 — Measured step response of delay network.

in the frequency range $| s | \ll \alpha$. Hence, in this limited frequency range the transfer function of the delay network becomes

$$\frac{V_4}{V_1}(s) \cong K - \left(1 + \frac{NR_2}{TR_1\alpha}\right) + \frac{NR_2}{TR_1\alpha} e^{-sT}, \tag{73}$$

which is a constant-delay, all-pass characteristic for

$$K = 1 + \frac{NR_1}{TR_1\alpha} = 1 + \frac{NR_2C}{T}. \tag{74}$$

An exact analysis of the circuit of Fig. 7 indicates that (73) is valid at low frequencies and that, by making the gain of the upper path, K, a frequency-dependent function, the constant-delay characteristic can be obtained over essentially the entire interval $| \omega | < N\omega_0/2$. The measured step response of the delay network is illustrated in Fig. 9. The N-capacitor element was constructed using a 64-contact rotary switch ($d/\tau \cong 0.61$) motor driven at a speed of 60 rps. Capacitors having a value of 0.1 microfarad were connected to each of the contacts.

A useful figure of merit for any delay network is its delay-bandwidth product. In this case, the delay is T seconds. The bandwidth is limited by that of the low-pass filter used to recover the output signal from the sampled data. This bandwidth cannot be greater than $1/2\tau$ cps, and $N\tau = T$, so that

$$\text{(delay) (bandwith)} \cong \frac{N}{2}. \tag{75}$$

5.2 *Narrow-Band Band-Pass Filter*

If the component networks in the N-path sampled-data networks have a low-pass characteristic with bandwidth small compared to ω_0, the transfer function, $F(0,j\omega)$, appears as a sequence of narrow passbands centered at multiples of ω_0, as previously noted. Consequently, this scheme is useful for the realization of highly selective band-pass filters. When only a single passband is required, the realization can be accomplished with a minimum value of $N = 3$, since the transfer function relation is valid for $|\omega| \leqq (N/2)\omega_0$. The band-limiting filter required at the output can also provide a low-frequency cutoff, so that the passband centered at zero frequency can be eliminated; the resulting transfer function is

$$T(j\omega) = \frac{1}{N}\left(\frac{d_1 d_2}{\tau^2}\right)[a_1 H(j\omega - j\omega_0) + a_1^* H(j\omega + j\omega_0)], \qquad (76)$$

where

$$a_1 = e^{j(\pi/T)(d_1 - d_2)} \left(\frac{\sin\dfrac{\pi d_1}{T}}{\dfrac{\pi d_1}{T}}\right)\left(\frac{\sin\dfrac{\pi d_2}{T}}{\dfrac{\pi d_2}{T}}\right).$$

This result is similar to the low-pass to band-pass transformation discussed in Section 3.1.

Since the band-pass characteristic is simply a frequency translation of a low-pass characteristic, it has arithmetic symmetry about the center frequency. Another advantage of this realization technique is that the filter can be easily tuned without altering the shape of the characteristic. The tuning is accomplished simply by changing the frequency of the timing source that controls the switching rate.

Implementation of the transfer function of (76) with series-sampling switches would require a current source at the input and negligible loading at the output. Analysis of the more practical circuit of Fig. 10, including source resistance, R_1, and load resistance, R_2, requires a somewhat different approach. Details of this analysis are carried out in Appendix B. The resulting transfer function is again a frequency-translated version of a low-pass characteristic:

$$T(j\omega) = \frac{E_2(j\omega)}{E_1(j\omega)} = \frac{Nd_2}{T}[a_1 G(j\omega - j\omega_0) + a_1^* G(j\omega + j\omega_0)]. \qquad (77)$$

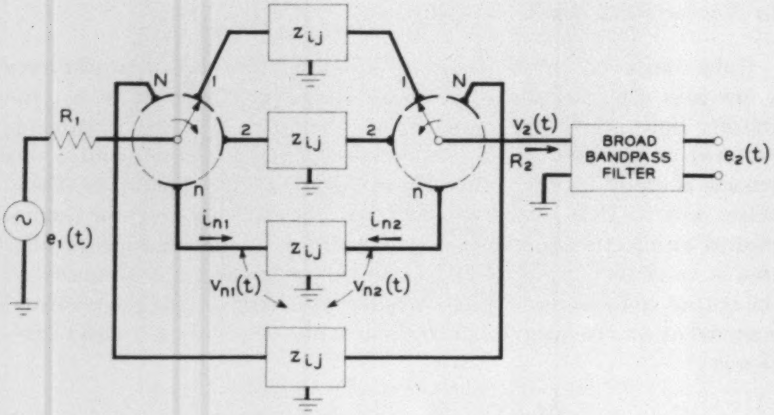


Fig. 10 — Practical circuit for realization of narrow-band filter characteristics.

The low-pass characteristic, $G(j\omega)$, is given by the voltage transfer ratio of one of the component networks operating between a source resistance, R_1T/d_1, and a load resistance, R_2T/d_2, as shown in Fig. 11.

A highly selective narrow-band filter using this scheme with $N = 4$ was constructed, using silicon diode input and output sampling switches controlled by two transistor multivibrator circuits. The center frequency of the filter was set at 25 kc. The low-pass component networks were three-section RC ladder networks with a bandwidth of approximately 3 cps. The Q-factor of a resonant circuit with the same bandwidth and center frequency would be greater than 4000. The selectivity of the sampled-data filter is even greater than the resonant circuit having this Q-factor, since the roll-off rate is greater. The measured frequency-response data and equivalent low-pass network are shown graphically in Fig. 12.

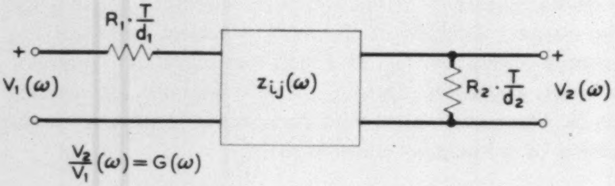


Fig. 11 — Equivalent low-pass network.

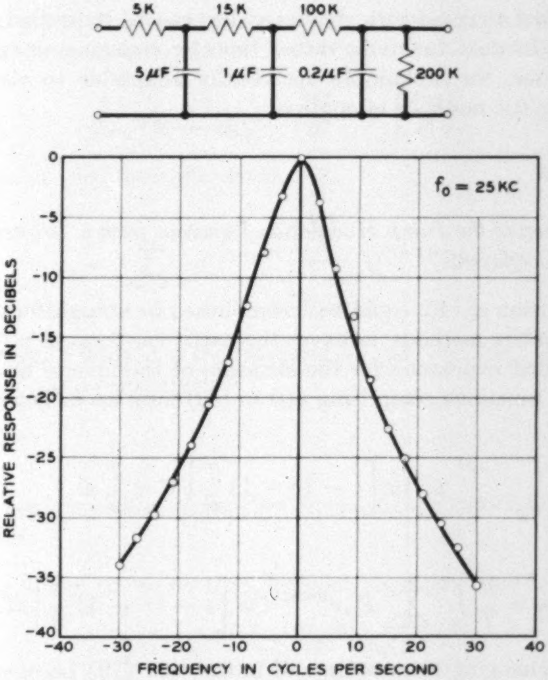


Fig. 12 — Frequency response and one of the constituent equivalent low-pass networks for sampled-data filter.

VI. CONCLUSION

The time-varying network configuration described in this paper exhibits several properties of both theoretical and practical significance.

A general input-output relation for the N-path structure has been derived. With the introduction of band-limiting restrictions, this relation can be expressed by a transfer function that is valid over a frequency band directly proportional to N, the number of parallel paths. In some special cases, however, band-limiting restrictions are unnecessary.

Several useful properties of the transfer function are maintained when the modulation is restricted to a type readily implemented by conventional switching techniques. The case where the modulators are replaced by series-sampling switches is examined in detail.

An important practical feature of the realization techniques discussed lies in the fact that network characteristics can be controlled by changing the modulation functions rather than by changing circuit element values. Hence, the techniques are readily adaptable to electronic or other automatic methods of control.

APPENDIX A

Determination of the Jump Modulation Function from a Prescribed Set of Fourier Coefficients.

The inversion of (40) could be accomplished by straightforward application of matrix methods, however the particular form of $p(t)$ affords a simple explicit expression for the elements of the inverse matrix. Note that the R functions comprising $p(t)$ in (40) form an orthogonal set, so that

$$\int_0^T p(t)w\left[t - (r-1)\frac{T}{R}\right] dt = \frac{T}{R} p_r. \tag{78}$$

Hence,

$$p_r = \frac{R}{T}\int_0^T \sum_{m=-\infty}^{\infty} P_m e^{j(m2\pi/T)t} w\left[t - (r-1)\frac{T}{R}\right] dt. \tag{79}$$

After interchanging summation and integration (79) becomes

$$\begin{aligned} p_r &= \frac{R}{T}\sum_{m=-\infty}^{\infty} P_m \int_{(r-1)(T/R)}^{r(T/R)} e^{j(m2\pi t/T)}\, dt \\ &= R\sum_{m=-\infty}^{\infty} P_m e^{-j(m\pi/R)} \frac{\sin\dfrac{m\pi}{R}}{m\pi} e^{j(2\pi/R)mr}. \end{aligned} \tag{80}$$

The values of P_m are not independent. From (42), it is seen that

$$P_{m+kR} = \frac{m}{m+kR} P_m. \tag{81}$$

Now (80) can be written as a finite sum over m:

$$p_r = R \sum_{m=-R/2}^{R/2}{}' P_m \frac{e^{-j(m\pi/R)} \sin\dfrac{m\pi}{R}}{\pi} e^{+j(2\pi/R)rm} \sum_{k=-\infty}^{\infty} \frac{m}{(m+kR)^2}, \tag{82}$$

where the prime on the summation over m is taken to mean that when R is even, the end terms of the series ($m = \pm R/2$) are added with half weight to avoid duplication in the sum over k. The series in k is summable and can be shown to be equal to

$$\sum_{k=-\infty}^{\infty} \frac{m}{(m + kR)^2} = \frac{m\pi^2}{R^2} \csc^2\left(\frac{m\pi}{R}\right). \tag{83}$$

Substitution of (83) into (82) gives

$$p_r = \sum_{m=-R/2}^{R/2}{}' e^{-j(m\pi/R)} \frac{\dfrac{m\pi}{R}}{\sin \dfrac{m\pi}{R}} e^{j(2\pi/R)rm} P_m . \tag{84}$$

This expression is written in the two forms of (43) and (44) for the cases of R odd and even respectively.

APPENDIX B

Equivalent Low-Pass Characteristic for Sampled-Data Realization of Band-Pass Filter

Referring to Fig. 10, we see that the following constraints are imposed:

$$\begin{aligned} i_{n1}(t) &= \frac{e_1(t) - v_{n1}(t)}{R_1} p_n(t), \\ i_{n2}(t) &= -\frac{v_{n2}(t)}{R_2} q_n(t); \end{aligned} \tag{85}$$

$$\begin{aligned} I_{n1}(j\omega) &= \frac{1}{R_1} \sum_{m=-\infty}^{\infty} P_m e^{-j(m2\pi/T)(n-1)\tau} \\ &\qquad \cdot [E_1(j\omega - jm\omega_0) - V_{n1}(j\omega - jm\omega_0)], \qquad (86) \\ I_{n2}(j\omega) &= \frac{1}{R_2} \sum_{m=-\infty}^{\infty} Q_m e^{-j(m2\pi/T)(n-1)\tau} V_{n2}(j\omega - jm\omega_0) . \end{aligned}$$

Representing the component networks in terms of open-circuit impedance parameters,

$$\begin{aligned} V_{n1}(j\omega) &= z_{11}(j\omega)I_{n1}(j\omega) + z_{12}(j\omega)I_{n2}(j\omega), \\ V_{n2}(j\omega) &= z_{21}(j\omega)I_{n1}(j\omega) + z_{22}(j\omega)I_{n2}(j\omega). \end{aligned} \tag{87}$$

Substitution of (86) into (87) results in infinite-order difference equations in $V_{n1}(j\omega)$ and $V_{n2}(j\omega)$ of the type normally encountered with periodically time-varying networks. However, the fact that the component networks are designed to have very narrow-band characteristics affords a considerable simplification of the equations.

If

$$|z_{ij}(j\omega)| \cong 0 \qquad \text{for} \quad |\omega| \geqq \frac{\omega_0}{2}, \tag{88}$$

then

$$\left.\begin{aligned} |V_{n1}(j\omega)| &\geqq 0 \\ |V_{n2}(j\omega)| &\geqq 0 \end{aligned}\right. \qquad \text{for} \quad |\omega| \geqq \frac{\omega_0}{2}. \tag{89}$$

The relations (88) and (89) permit the elimination of all terms except $m = 0$ in the sums involving $V_{n1}(j\omega)$ and $V_{n2}(j\omega)$. Hence, for $|\omega| \leqq \omega_0/2$,

$$\begin{aligned} &\left(1 + \frac{z_{11}}{R_1} P_0\right) V_{n1} + \frac{z_{12}}{R_2} Q_0 V_{n2} = \\ &\qquad\qquad \frac{z_{11}}{R_1} \sum_m P_m e^{-j(m2\pi/T)(n-1)\tau} E_1(j\omega - jm\omega_0), \end{aligned} \tag{90}$$

$$\begin{aligned} &\frac{z_{21}}{R_1} P_0 V_{n1} + \left(1 + \frac{z_{22}}{R_2} Q_0\right) V_{n2} = \\ &\qquad\qquad \frac{z_{21}}{R_1} \sum_m P_m e^{-j(m2\pi/T)(n-1)\tau} E_1(j\omega - jm\omega_0), \end{aligned} \tag{91}$$

where

$$P_0 = \frac{d_1}{T}, \qquad Q_0 = \frac{d_2}{T}.$$

Eliminating V_{n1} from (90),

$$V_{n2}(j\omega) = \frac{T}{d_1} G(j\omega) \sum_m P_m e^{-j(m2\pi/T)(n-1)\tau} E_1(j\omega - jm\omega_0), \tag{92}$$

where

$$G(j\omega) = \frac{z_{21} R \dfrac{T}{d_2}}{\left(z_{11} + \dfrac{R_1 T}{d_1}\right)\left(z_{22} + \dfrac{R_2 T}{d_2}\right) - z_{12} z_{21}}. \tag{93}$$

The output voltage of the N-path configuration is given by

$$
\begin{aligned}
v(t) &= \sum_{n=1}^{N} v_n(t) q_n(t), \\
V_2(j\omega) &= \sum_{n=1}^{N} \sum_{l=-\infty}^{\infty} Q_l e^{-j(l2\pi/T)(n-1)\tau} V_{n2}(j\omega - jl\omega_0).
\end{aligned}
\tag{94}
$$

Substituting (92) into (94),

$$
\begin{aligned}
V_2(j\omega) = \frac{T}{d_1} \sum_{n=1}^{N} \sum_{l,m} Q_l P_m e^{-j[(l+m)2\pi(n-1)/T]\tau} \\
\cdot G(j\omega - jl\omega_0) E_1[j\omega - j(l+m)\omega_0]
\end{aligned}
\tag{95}
$$

and, summing over n as was done in (9),

$$
V_2(j\omega) = \frac{NT}{d_1} \sum_l Q_l G(j\omega - jl\omega_0) \sum_k P_{k-l} E_1(j\omega - jkN\omega_0). \tag{96}
$$

Now, if $E_1(j\omega)$ is band-limited such that

$$
|E_1(j\omega)| \cong 0 \quad \text{for} \quad |\omega| \leqq \frac{N\omega_0}{2} \tag{97}
$$

and $V_2(j\omega)$ is followed by a low-pass filter with cutoff at $\omega = N\omega_0/2$,

$$
\frac{E_2(j\omega)}{E_1(j\omega)} = \frac{NT}{d_1} \sum_l Q_l P_{-l} G(j\omega - jl\omega_0). \tag{98}
$$

Now, suppose that the low-pass filter is replaced by a band-pass filter that selects only the passband corresponding to $l = \pm 1$. Then,

$$
\frac{E_2}{E_1}(j\omega) = \frac{Nd_2}{T}[a_1 G(j\omega - j\omega_0) + a_1^* G(j\omega + j\omega_0)], \tag{99}
$$

where

$$
a_1 = e^{j(\pi/T)(d_1 - d_2)} \left(\frac{\sin \frac{\pi d_1}{T}}{\frac{\pi d_1}{T}} \right) \left(\frac{\sin \frac{\pi d_2}{T}}{\frac{\pi d_2}{T}} \right).
$$

The transfer function of (99) is equivalent to that of (76), where the low-pass function, $G(j\omega)$ in this case, is simply related to the low-pass characteristic of one of the component networks. Examination of the relation (93) shows that $G(j\omega)$ is simply the voltage transfer ratio of one of the component networks operating between a source resistance of $R_1(T/d_1)$ and a load resistance $R_2(T/d_2)$, as shown in Fig. 11. This equivalent low-pass network provides the basis for synthesis of the prescribed band-pass characteristic.

REFERENCES

1. Paris, H., Utilization of the Quadrature Functions as a Unique Approach to Electronic Filter Design, I.R.E. Conv. Rec., 1960.
2. Linvill, W. K., The Use of Sampled Functions for Time-Domain Synthesis, Proc. Nat. Electronics Conf., Chicago, September 29–30, 1953, Vol. 9, p. 533.
3. LePage, W. R., Cahn, C. R. and Brown, J. S., Analysis of a Comb Filter Using Synchronously Commutated Capacitors, A.I.E.E. Trans., Part I, **72**, 1953, p. 63.
4. Smith, B. D., Analysis of Commutated Networks, I.R.E. Trans., **PGAE-10**, p. 21.

Magnetic Latching Crossbar Switches:

A New Development in Magnetic Properties of Tool Steel

By F. A. ZUPA

(Manuscript received April 12, 1960)

A magnetic latching function in crossbar switch hold magnets is obtained by means of a specially designed magnet core made of high-carbon tool steel. The fabricated core detail is given a hardening heat-treating cycle, regulated to produce a particular degree of physical hardness that was found to impart the optimum combination of magnetic properties needed to obtain pulse operation and magnetic latching of the electromagnet under a wide range of contact spring loads. The nominal latching force developed with this new electromagnet design is ¼ lbs, with a cylindrical core of only 0.11 square inch cross-sectional area. The electrical operating power need be only 2.5 watts applied for 0.100 second or about 18.0 watts for 0.015 second, and the reverse release pulse strength is about 50 per cent of the operate value. The coexisting values of coercive force, residual induction and magnetic permeabilities obtained in this design are new and useful to the art of designing electromagnetic switching devices with a magnetic latching function.

I. INTRODUCTION

Since the introduction of the dial-type telephone switching systems, switching devices such as relays and electromagnets have become the most essential and widely used of all the components in the telephone central office. Many notable improvements on these switching devices have made it possible for the telephone systems to grow and serve the increasing population of customers. The advancements on these devices have dealt largely with their sensitivity and speed of operation, contact switching capacity, service life and reliability. In contrast to these improvements, however, it appears that very little has been done to save operating power by utilizing residual magnetic energy to effectively hold the electromagnets in the operated position without continuous current drain. Recently, however, a new electromagnet core design that provides this function was developed for crossbar switch hold magnets. This utilizes a new combination of magnetic properties that have been found to exist in high carbon steel after it has undergone a suitable hardening heat-treating cycle.

There are no mechanical locking features associated with this new magnetic latching hold-magnet design. The magnetic latching force developed at the termination of the short electrical operating pulse is obtained solely by the efficient use of the residual magnetic induction and coercive force properties of the new magnet core. To restore the electromagnet to its nonoperated position, it is only necessary to re-energize the magnet coil with another short pulse of lower current strength and opposite polarity.

The total amount of electrical power necessary to energize the magnetic latching hold magnet is about 2.5 watts applied for only 0.100 second. Since many hold magnets must hold during each telephone conversation, this represents a very large power saving compared to the power used by the present nonlatching hold magnets. This design of magnetic latching hold magnets makes it possible to use 100- and 200-crosspoint crossbar switches in remote locations where the power supply is very small compared to that in a central office. A notable application of this new magnet core development is the conversion of existing crossbar switch hold magnets to magnetic latching operation, as might be used in telephone line concentrators.

II. NEED FOR A NEW MAGNET CORE MATERIAL

The state of the art in the design of electromagnets and the processing of associated magnetic materials for useful magnetic properties has advanced with many notable improvements during the past thirty years. It is of interest to observe the direction that some of the improvements in magnetic materials have taken in relation to what is required for magnetic latching functions.

In the class of soft magnetic materials, such as the magnetic irons and low carbon steels normally used for relays and electromagnets, the effort has been directed mainly toward greater permeability and associated reduction of coercive force. Since this is in the direction of reducing the quantity of the stored electromagnetic energy, usually represented by the product of the coercive force and remanence, this class of materials is definitely not suitable for a magnetic latching function. The property of low coercive force and associated greater permeability, of course, is very useful for obtaining greater operating sensitivity and greater release-to-operate ratios.

In the class of hard magnetic materials normally used for permanent magnets, the effort has been directed to increase the coercive force, even at the expense of a reduction in remanence, as long as the result was an increase in the numerical value of the product of coercive force in oersteds and remanence in gausses. In spite of the high values of magnetic energy that can be stored in them, permanent magnet materials would not be satisfactory in the core of a magnetic latching electromagnet, primarily because the required operating power would be several times as high as is practical in switching circuits. In general, this is due to the inherent high magnetic reluctance or low magnetic permeability of hard magnetic materials that are processed to be permanent magnets. Magnet cores that are made from materials commonly used for permanent magnets are therefore not conducive to efficient magnetic latching designs, especially when the contact spring loads on the same electromagnet range from small to large values from one operation to another, as they do in crossbar switch hold magnets. The required range of contact spring loads will be described later.

It appears, therefore, that past developments in magnetic materials have not been in the direction of producing a high order of quality in both operating and magnetic latching properties. The development of an economical and workable magnetic latching hold magnet design required the development of new coexisting combination of values of permeability, coercive force and remanence in a suitable magnet core material. A description of this development and the resulting operating capabilities of the magnetic latching crossbar switches that have been designed for new telephone equipment will be given, with special emphasis on the essential electromagnet design principles that guided this development.

III. BASIC FACTORS GOVERNING DESIGN OF THE LATCHING MAGNET

The combination of magnetic properties that must be obtained in the magnetic circuit of the electromagnet to satisfactorily meet the operating and latching functions is dependent upon the following primary factors:

i. the permissible mechanical form and size of the electromagnet and its switching functions;

ii. the range of contact spring loads to be applied to one magnet assembly;

iii. the range of the electrical pulses, in time and power values, available to operate and release the magnet.

It is therefore desirable to first describe these conditions, in order to follow the steps taken in the magnet core development.

3.1 *The Structure and Switching Functions of the Crossbar Switch Hold Magnet*

The magnetic latching hold-magnet design will be used for the same type of contact switching functions as those of the nonlatching hold magnets presently used in the crossbar switches of crossbar switching telephone systems, except that the loads will cover a wider range of values. As illustrated by Fig. 1, the hold magnet is the motor element of the vertical unit assembly. The latter, as its name implies, provides a vertical row of ten levels of crosspoint contacts, each level consisting of two to six pairs of make contact springs that are used for transmission and control circuit connections, and a separate assembly of hold-off normal contact springs (HON), consisting of two or three pairs of make or break contacts that are used for common control circuit connections.

The select magnets and vertical units are mechanically linked by horizontal select bars carrying flexible wire fingers that can be rotated through a small angle in either of two directions. The crossbar switch therefore represents a rectangular coordinate arrangement of 100 or 200 crosspoints, any one of which may be selected by the operation of

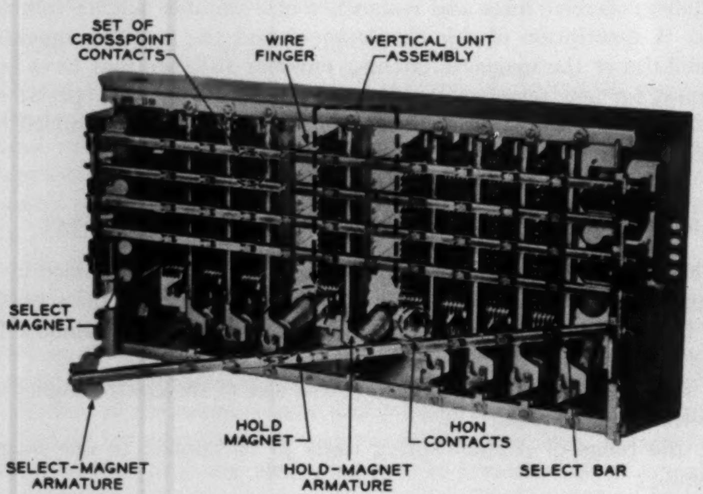


Fig. 1 — Crossbar switch, showing location of hold magnet as motor element of vertical unit assembly.

a particular select magnet and hold magnet. The total number of contacts that may be actuated by one hold magnet depends upon the circuit operating sequence of the select magnets in the crossbar switch, as described below.

The operation of a select magnet rotates the select bar associated with it, thereby interposing a wire finger between each hold magnet armature and the end of each card supporting the moving springs of each set of crosspoint contacts that lies in the horizontal level corresponding to the operated select magnet. Then the operation of a particular hold magnet determines which set of crosspoint contacts is selected in that horizontal level. Sometimes two select magnets are energized simultaneously in order to select one set of crosspoint contacts in each of two horizontal levels by the operation of one hold magnet. Sometimes the hold magnet is operated to switch the HON contacts without any crosspoint contacts. The quantitative values of the different contact spring loads that may be applied to one hold magnet are shown graphically in Fig. 2.

3.2 *Mechanical Load Forces Affecting the Design of the Magnet*

Each curve of Fig. 2 shows the rate at which the spring load builds up on the hold magnet armature, as the armature moves from the nonoperated position to the operated position against the core poleface. As indicated, the maximum crosspoint and HON contact spring load may build up to a value of 1150 grams and the minimum HON spring load may be only 140 grams. These individual load values are very important, because the new hold magnet design, to be successful, must be capable of operating, latching and releasing with any one of the load values, under any one of the extremes of the circuit operating power conditions.

There are two important magnetic requirements on the new magnet design that are affected by the maximum load build-up rate shown in Fig. 2. The first is that the magnetic force of attraction acting on the armature during its operating travel shall always exceed the force required to move the corresponding instantaneous load by a substantial amount. It is this differential, together with the electrical time constant of the magnet coil (the time rate of coil current development), that governs the operating or switching time of the electromagnet. The second requirement on the core is that the magnetic latching force shall always exceed by a substantial amount the force required to hold the maximum load of 1150 grams. It is this differential that governs the ability of the latched magnet to withstand disturbing forces that

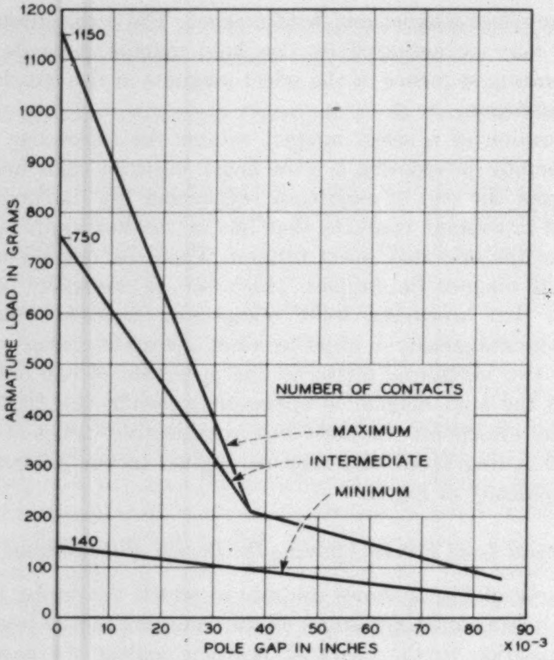


Fig. 2 — Range of contact spring loads on magnetic latching hold magnets.

may be developed by shock and vibration when the equipment is mounted on a telephone pole. If the disturbing vibrations should cause the armature to bounce or lift off the core poleface by only a fraction of one mill-inch, the spring load might then cause the premature release of the armature. More will be said later about the latching force margin, the disturbing forces and the effect of very small separations between the mating poleface surfaces.

The minimum load value of 140 grams is also an important consideration in the new magnet design, because it affects the permissible limits of the strength of the reverse release pulse that may be applied to the latched electromagnet without false reoperation. This means that the minimum electrical strength of the release pulse must be strong enough to always release the lightly loaded armature, but that the maximum pulse strength must not reoperate it. Failure to release or false reoperation are trouble conditions that must be guarded against in the magnetic latching design.

3.3 *Operate and Release Pulses Prescribed by the Circuit Conditions*

Since the electrical power available in some of the circuits that will use the new switches is limited, it was necessary to place a limit on the maximum current strength in the individual operating pulse for the magnetic latching magnet design. Circuit and equipment design considerations also determined the limiting values of the voltage and the time duration of the operating and releasing pulses. The limiting pulse values, insofar as they affect the magnet design, were set up, tentatively, to be as follows:

Energizing pulse: maximum 0.2 ampere at 22 to 28 volts for a minimum of 0.100 second;

Operating time (to switch all contacts): maximum 0.050 second;

Releasing time (to restore all contacts): maximum 0.050 second.

IV. DEVELOPMENT OF THE MAGNET CORE DESIGN

From the foregoing analysis of the work loads and the power available to perform the electrical operate, magnetic latching and unlatching functions, the level of the magnetic properties that should be available in the magnetic circuit of the new electromagnet design can be estimated. It should be noted also that, while the magnetic circuit consists of a core, an armature and a yoke or return polepiece, in order to maintain the present construction and mode of operation of the crossbar switch, the first efforts were directed to realize the design objectives with only a simple change in the material and design of the core.

The next step taken in the development of the design, therefore, was to make an analysis of the commercially available magnetic materials that might be suitable for the new magnet core design. Since the maximum contact spring loads represented by Fig. 2 are comparable to those of the present nonlatching hold magnets, the new magnet core material had to be capable of developing a level of magnetic induction strength that was not much below that of the presently used core, which is made of annealed low-carbon steel, in order to operate the electromagnet on reasonable values of magnetomotive force. The residual magnetic induction of the material, however, should be supported by a much stronger coercive force value, in order to produce and maintain the desired high level of magnetic latching force. It appeared that one type of magnetic material that should be considered was the magnet steels, which can be processed to develop (a) high flux strength at reasonably low magnetizing forces and (b) high remanence with a suitable value of coercive force. A brief analysis of the essential magnetic properties of known steels that have at least some of the desired magnetic characteristics is given below.

Table I shows a comparison of the pertinent magnetic properties of (a) annealed 0.10 per cent carbon steel, which is widely used for the magnetic circuit of many types of electromagnetic switching devices with simple operating requirements; (b) hardened 0.9 per cent carbon steel and (c) hardened 5 per cent tungsten steel, both of which were used for making permanent magnets about 50 years ago, before the development of more efficient permanent magnet alloys containing less iron and more of other alloying elements.

The 0.10 carbon steel has adequate values of magnetic permeability and saturation induction to develop the required open-polegap tractive forces. Its coercive force, however, is too low to retain the residual flux density required to produce the needed latching force. The hardened high-carbon and tungsten steels have the necessary coercive force, but their permeability is too low to develop the required values of flux densities with the available operating power. The combination of values of magnetic properties needed to develop the required tractive and latching forces, with the operating magnetizing force available in the electromagnet, will be described later.

4.1 *Selection of Magnet Core Material for Study*

It appeared, therefore, that the magnetic properties required to meet the desired operating and latching functions were in between those of the annealed low carbon steel and the hard permanent magnet type of steel. Since the magnetic properties of high carbon steels are known to vary with the hardness of the physical structure of the steel, it was conceived that a critical study of this relation, instead of the usual

TABLE I—TYPICAL DATA FOR ANNEALED LOW-CARBON MAGNET STEEL AND HARDENED HIGH-CARBON PERMANENT MAGNET STEEL

Magnetic Characteristic	Annealed 0.10 Carbon Steel	Quench-Hardened 0.9 Carbon Steel	Quench-Hardened 0.7 Carbon 5.0 Tungsten Steel
Saturation induction, B_s, in gausses	21,000	12,000	13,000
Residual induction, B_r, in gausses	10,000 to 14,000	8,500 to 10,000	8,500 to 10,300
Coercive force, H_c, in oersteds	1.8	50	70
Permeability, μ_{max}	2,000	111	123

Note: These data are representative of the magnetic properties obtained with test ring samples of the material and the magnetizing force (H_{max}) value is generally 300 or more oersteds.

study relating magnetic properties to heat-treating temperature cycles, should disclose the best combination of operating and latching magnetic properties possible with the high carbon steel. This new approach to evaluate the magnetic properties of hardenable steel was better than relying only on the measured temperatures and time of the heat-treating cycles, because the iron-carbon alloys resulting from the latter cycles usually vary considerably with the size and shape of the specimens. A laboratory study was therefore undertaken to determine the quantitative relation between the measured physical hardness produced by controlled heat treatments and the magnetic operating and latching properties, using a commercially available high-carbon steel for the magnet core test specimens.

In order to carry out the above study so that the results would be directly applicable to the magnet core design, the type of high-carbon steel selected for the study was determined on its merits from the standpoint of uniformity in composition and commercial availability in the round stock size best suited to the hold magnet design, 0.375 inch diameter. With these factors in mind, a tool steel having the nominal composition of iron plus 1.2 per cent carbon, 0.3 manganese, 0.22 silicon, 0.10 vanadium, 0.025 sulfur and phosphorous was selected. This grade of steel has been used for many years by the machine industries, primarily for making hardened tools and machine parts. Machine shop practices on the quenching and tempering of parts made from this grade of tool steel show that the parts can be hardened over a wide range of hardness values by first heating them to about 1475°F, immediately quenching in a liquid cooling medium (water or oil), then reheating at a lower temperature and slowly cooling in air at room temperature, the value of the reheating temperature being the principal determinant of the physical hardness of the parts. It should be noted, however, that the time cycles of heating and cooling, and the ambient atmospheric conditions during heating from the standpoint of minimizing decarburization, have important effects on the resulting chemical and physical changes that take place in the structure of the steel parts. The laboratory study therefore was planned with well-controlled experiments in heat treatment and the evaluation of the associated magnetic properties that control the operate, latching and unlatching functions in the electromagnet.

4.2 *Development of the Magnet Core Poleface Design*

In order to have the results of the experiments on the magnetic properties of the steel specimens directly applicable to the hold-magnet design, the size and shape of the test specimens were designed to represent an efficient magnet-core design. It is of interest, therefore, to examine the effect of the size and shape of the core poleface surface on the operating and latching characteristics of the electromagnet. The importance of the poleface design cannot be overemphasized, because the working margins obtained in the operating and latching capabilities of the magnet are largely affected by the poleface design. Some of the functional aspects of the poleface design are discussed below. In this discussion the core specimen is assumed to be a $\frac{3}{8}$-inch-diameter rod, approximately 3.5 inches in over-all length, because this is the maximum size that can be conveniently used in the present crossbar switch structure.

The following general relation between poleface area and magnetic force of attraction may be used to estimate the optimum value of the area for (a) the open polegap force and (b) the closed polegap or latching force:

$$F = \frac{\Phi^2}{8\pi A}\left(\frac{1}{980}\right)k$$

where F = the force in grams,

Φ = the magnetic flux in maxwells, between the poleface area A and the mating surface area on the armature,

A = the poleface area in square centimeters,

k = a constant, the value of which corrects for the nonperpendicularity in the direction of Φ between the mating polefaces.

Based on experience with flux measurements on this type of magnetic circuit design, the value of k is slightly less than one for the closed polegap condition. For the open polegap conditions, the greater the gap the smaller is that value.

Since the value of the polegap flux Φ is determined by the applied coil ampere-turns and the corresponding values of magnetic reluctances prevailing in the complete magnetic circuit of the electromagnet, one important portion of which is that of polegap, the general effect of poleface area A on the force F can be described by referring to the ampere-turn and reluctance form of the force equation

$$F = \frac{2\pi(\mathrm{NI})^2}{A(R_0 + R_g)^2}$$

where NI = ampere-turns,

R_0 = sum of all reluctances in the magnetic circuit except that of the polegap,

$R_g = l/\mu A$ = polegap reluctance,
l = length of the polegap,
μ = permeability of air and metal finishes in the polegap.

It can be shown, therefore, that when l is very small, as it is in the latched condition of the polegap, since R_g is then also relatively small, the value of F is made greater by using a smaller value of A up to the limit when its value results in a significant increase in the value of $R_0 + R_g$.

Conversely, when the values of l are relatively large, as they are during the operating travel of the magnet armature, the corresponding values of R_g are large enough to be controlling in their effect on the values of F. Then the value of F is made greater by increasing the area A up to the limit when its effect on the value of $(R_0 + R_g)^2$ is no longer significant.

Another important consideration in the design of the core poleface was its shape or geometry. This factor deals with the uniformity of the closed-polegap reluctance, as affected by the relative alignment of the armature poleface surface against that of the core. It is well known that two mating flat poleface surfaces usually make only a line contact and therefore result in an angular airgap. In this magnet design, a forward displacement of about 0.005 inch in the position of the core with a plane poleface would result in a separation of about 0.003 inch at the center of the core poleface. To avoid the detrimental effect of unavoidable misalignments, the poleface surface on the core was shaped like the surface of a 16-inch-diameter sphere, while the mating poleface surface on the armature was flat (commercial quality). As can be seen from the sketches in Fig. 3, the common contact area between a flat and a spherical surface is affected comparatively little when the core is displaced about 0.005 inch. Under common manufacturing conditions, therefore, the use of a large-radius spherical poleface mating with a flat poleface results in considerably less variation in the closed polegap reluctance, particularly with the type of hold magnet structure shown in Fig. 4.

This is by no means intended to represent a complete discussion of the effects of poleface area and shape on the magnetic force of attraction. It is sufficient to show, however, that the latching force is greater with a smaller poleface area at the expense of some loss in the force of attraction at the large open polegaps, and that the strength and uniformity of the latching force are better with the spherical surface.

Referring to Fig. 4, observe that the magnet core of the nonlatching design (present crossbar switches) has a poleface area much larger than the cross-sectional area of the core, the enlarged poleface being produced by an automatic cold heading operation on the soft steel rod. This poleface design was made to obtain greater efficiency in producing the open polegap tractive forces. The magnet core of the magnetic latching design (new crossbar switches for the line concentrator), however, requires a much smaller poleface area. The value of its area was determined on the basis of providing a latching force of at least 1450 grams, in order to have about 25 per cent margin above the latching force required to hold the maximum load of 1150 grams. This margin was determined by estimating the effect of vibrations and shocks on the hold magnet when the crossbar switches and associated equipment are mounted on a telephone pole. Available data on the amplitudes and frequencies of vibration that may occur on a telephone pole indicated that the resulting acceleration may be as high as 1 g at the mounting position of the crossbar switch. In view of the wide range of compliances and masses in the structural parts of the crossbar switches, the estimated margin of minimum 300 grams between the maximum load and the minimum latching force was considered a suitable temporary value, until confirmed by laboratory vibration tests.

In order to develop a minimum latching force of 1450 grams with a

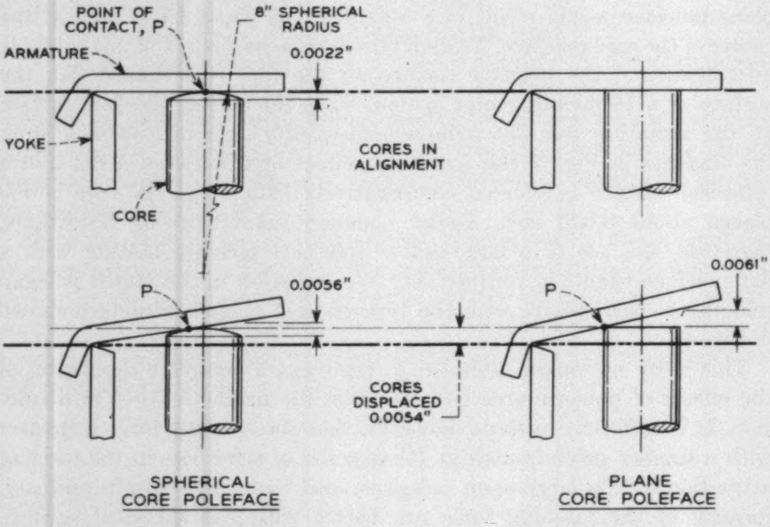


Fig. 3 — Schematic showing general effect of armature misalignment on closed-polegap reluctance.

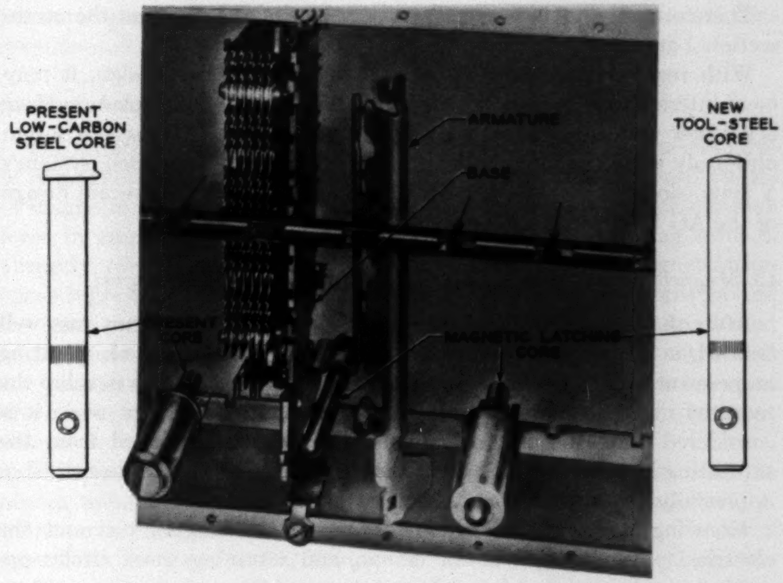


Fig. 4 — Hold magnet structure with large-radius spherical poleface mating with flat poleface.

$\frac{3}{8}$-inch-diameter magnet core, the most efficient poleface area was estimated by assuming expected values of residual magnetic induction and coercive force in the body of the prospective magnet core, under the influence of the self-demagnetizing action of the magnetic circuit. As shown by the B_r values given in Table I, it appeared reasonable to assume that the residual induction of a high quality steel core should be at least 9,000 gausses. With these assumed values of latching force and residual flux density the estimated poleface area A was obtained as follows:

In the relation

$$F = \frac{\Phi^2}{8\pi A}\left(\frac{1}{980}\right)k$$

let $F = 1650$ grams (average value) and $k = 0.85$ (estimated value).

Since the cross-sectional area of the core is 0.71 sq cm, a flux density of 9000 gausses represents 6300 maxwells in the core. Assuming a loss of 10 per cent due to core surface leakage, the value of Φ reaching the poleface is 5670 maxwells.

Therefore, $A = 0.651$ sq cm, or very nearly the same as the cross-sectional area of the core.

With regard to the importance attached to poleface design, it may be of interest to note that the mating of a spherical with a plane surface is not new with this magnetic latching design. It was first used to obtain uniformly closed polegap reluctances in the design of the Bell System's Y type (slow release) relay, in 1935, and again in the more recent design of the AG type (slow release) relay.

4.3 *Magnetomotive Force Available To Energize the Electromagnet*

After the structural size and shape of the magnetic circuit was well defined, it was necessary to determine the minimum level of operating ampere-turns that would be available in the magnet coil to develop the required magnetic properties. The need for this is apparent when it is considered that the residual magnetic properties obtained from the saturating level of magnetization are different than those obtained from apprecially lower levels.

Knowing the available winding space in the magnet coil and the electrical pulse strength in the circuit, and assuming worst circuit operating conditions under outdoor extreme temperatures of −40° to +140°F, the steady-state value of coil ampere-turns available to energize the electromagnet was found to be a minimum of 565 and a maximum of 1065. This wide range of magnetomotive force was partly due to a circuit condition that placed two of the magnet coils in parallel and both in series with a protective lamp. These extremes of circuit operating values account for the importance attached to the minimum and maximum total load values described earlier.

V. INVESTIGATION OF MAGNETIC PROPERTIES WITH THE NEW MAGNET CORE DESIGN

The purpose of this investigation was to determine whether the selected high-carbon tool steel core could be made to yield a combination of pertinent magnetic property values when the magnet was energized with the available magnetomotive force values. The processing of the steel core, of course, was to be a reproducible hardening heat treatment. The essential experiments and test results in this investigation can now be described in relation to the desired design capabilities. Since the purpose of this study was to find the relation between the physical hardness of the steel core, as produced by hardening heat treatments, and the resulting pulse operating and latching characteristics, a practical test method was used to determine the relation, in addition to the direct measurements for magnetic characteristics of the individual steel core specimens.

5.1 *Procedure for Evaluation of Test Results*

The criterion used for the appraisal of the test results is a special form of demagnetization curve plotted in terms of the instantaneous values of magnetic latching force in grams and demagnetizing magnetomotive force in ampere-turns, as the applied saturation magnetizing force is abruptly reversed to the demagnetizing value. Each demagnetization curve represents the typical data obtained on several test cores having the same particular level of physical hardness, and each core was tested with the same hold magnet structure and coil. It should be noted that the preparation of the test-core specimens involved the establishment of uniform machining of the core poleface and uniform heat treatment processes, in order to minimize extraneous variables.

With regard to the determination of physical hardness, in order to obtain data directly applicable to subsequent manufacturing test requirements, each test core was measured on the 30-N scale of a Rockwell superficial hardness tester, before the corrosion protective finish was applied to its surface. This nondestructive and simple method of measuring hardness is one of the accepted inspection testing methods. However, since it measures hardness to a depth that is only a small fraction of the cross section, its accuracy depends upon the uniformity of the hardness throughout the volume of the test specimen. This presented no serious problem, because the small radial depth and uniform section of the core specimens assures a reasonably uniform hardness.

With regard to the Rockwell hardness numbers used to designate the physical hardness of each test specimen, it should be noted that they represent the actual 30-N scale readings as taken on the cylindrical surface of the 0.375-inch-diameter cores before the application of the protective finish. In order to reproduce the same physical hardness represented by these Rockwell hardness numbers on parts having different radii of curvature or having flat surfaces, the numerical values should be corrected according to the empirical tables furnished with the Rockwell tester. For example, in our data, the hardness readings from 54 to 64 would become 55.5 to 65 when converted to represent readings on flat surfaces.

5.2 *Magnetic Latching Forces Versus Hardness*

The characteristics of two cores of widely different degrees of hardness are shown in Fig. 5, one representing the maximum and the other

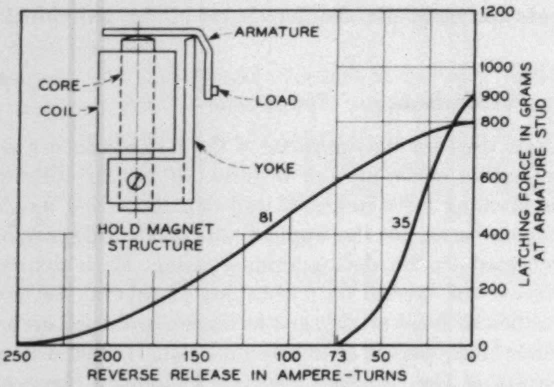


Fig. 5 — Magnetic latching characteristics of hold magnets with cores of maximum and minimum physical hardness.

the minimum hardness. Each curve is designated by the Rockwell 30-N hardness number, as measured on the cylindrical surface of the test core. The number 81 represents the maximum hardness value, as obtained after quenching and then reheating to a stress-relieving temperature of 350°F; the number 35 represents the minimum hardness value on the core, as obtained with a high-temperature (about 1600°F) normalizing heat treatment. Observe that the number 81 (hard) core developed an open circuit latching force of 800 grams, which is only 57 per cent of the required minimum value. Its demagnetizing pulse strength, however, was 250 NI, a value that is greater than desired for controlling the release of the electromagnet with the minimum load of 140 grams. In contrast to this permanent-magnet type of core, observe that the number 35 (soft) core developed a latching force of 900 grams, while its demagnetizing value was only 73 NI. It was evident, therefore, that neither of these cores representing extreme levels of physical hardness had the necessary magnetic residual induction strength. More details on their magnetic characteristics will be given later, by showing some of the actual magnetization hysteresis loops of the test-core specimens.

Fig. 6 shows the magnetic latching characteristic curves of the two cores having Rockwell hardness numbers of 72 and 41, together with the former set of curves for comparison. Observe that the number 72 hardness core developed a latching force of 1200 grams, while the number 41 core developed a latching force of 1350 grams. Compared to the slightly harder and softer cores with hardness numbers of 81 and 35, respectively, a gain of 50 per cent in latching force is realized for each

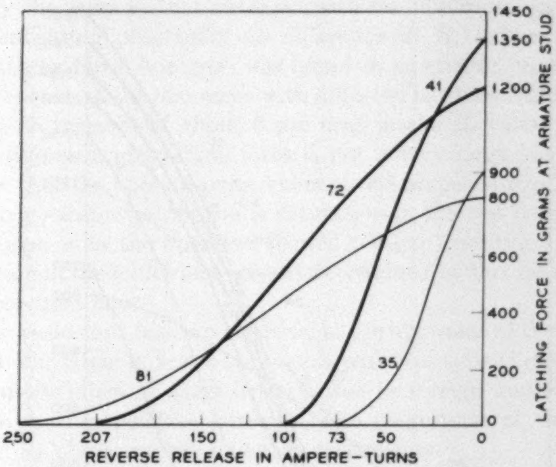


Fig. 6 — Effect of small changes in hardness of core on resulting latching force.

of the two intermediate hardness cores. Observe also that the corresponding demagnetizing ampere-turn values have changed considerably, the more important change being on the number 41 hardness core.

From the designer's viewpoint, the above results are very encouraging, in spite of the fact that the open-circuit latching force is still appreciably below the required minimum of 1450 grams. It is significant that a 50 per cent increase in latching force results from a relatively small change in physical hardness. The rate at which this improvement is made by the remainder of the intermediate hardness values is therefore of even greater interest.

Fig. 7 shows an additional set of four characteristics curves, each representing a different level of core hardness, and this completes the range of core hardness levels that was investigated. Examination of the added curves shows that the open-circuit latching force continues to increase from both ends of the hardness range, and that the optimum latching force value of 1800 grams occurs with the number 60 hardness core. The demagnetizing ampere-turn value, however, continually decreases as the hardness number decreases.

The eight magnetic latching characteristic curves in Fig. 7 show that there is an outstanding improvement in the magnetic properties of the tool steel cores when their physical hardness, as produced by hardening heat treatments, is in the Rockwell hardness range (30-N) of 54 to 64.

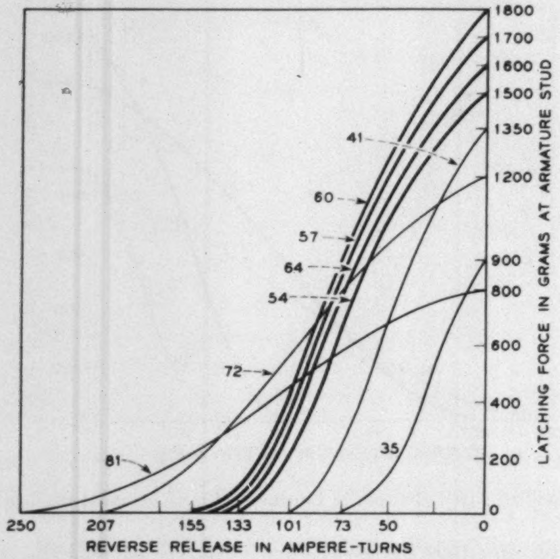


Fig. 7 — Magnetic latching characteristics of additional hold magnets with cores of different physical hardness.

The latching force values of 1500 to 1800 grams that were obtained in this hardness range provide the working margins that are necessary for the crossbar switch hold magnets. It is of interest, therefore, to examine these results from the standpoint of their reproducibility and associated variables. Also, it is desirable to examine the basic magnetic properties of the core material with these hardening heat treatments, in terms of values that can be used for other possible design applications.

The electrical operate-soak value, which determined the level of magnetic flux density established in each test core prior to the measurements for its latching force and reverse release characteristic, was kept constant at the minimum worst circuit pulse value of 565 ampere-turns. It should be noted that, with operate-soak values of greater magnetizing force, the latching characteristics are slightly different, because the resulting residual induction (B_r) values tend to be greater, while the coercive force (H_c) values are not appreciably different. The magnitude of these effects is indicated by the following test results obtained with the same test cores.

With an operate value of 960 ampere-turns, the reverse-release ampere-turn value required to reduce the latching force to zero was found to be practically the same as that obtained with the minimum operate value, thereby indicating practically no difference in H_c values. The open-circuit latching force, however, was found to be greater, up to about 10 per cent for each of the test cores with different hardness values, thereby indicating an increase of about 5 per cent in the B_r value. Since a 10 per cent difference in latching force is not a very large increase, these data show that the operate pulse value of 565 ampere-turns is sufficient to develop a substantial magnetic saturation in the test cores when the electromagnet is in the operated (closed polegap) position. The degree of saturation in the individual core, as determined by flux measurements, will be presented later.

Another important factor considered in the appraisal of the magnitude of open-circuit latching forces obtainable with this type of electromagnet design was the effect of small irregularities or foreign matter on or between the mating poleface surfaces. The magnitude of this effect is illustrated by observing the following test results.

With a given core of optimum magnetic properties (core with number 60 hardness value) assembled in a normal electromagnet, and the armature and core poleface surfaces being of good commercial smoothness and coated with a commercial nickel protective finish, the introduction of a 0.0005-inch-thick nonmagnetic separator between the mating polefaces was found to reduce the open circuit latching force by as much as 15 per cent. The reason for this effect, it can be shown, is that the added 0.0005-inch airgap increases both the flux leakage and the magnetic reluctance at the closed polegap. Since the latching force varies directly as the square of the flux value, a loss of about 7 per cent in the effective residual flux would account for a loss of about 15 per cent in force. It is obvious, therefore, that a protective finish of nickel (due to its magnetic permeability) is more desirable than a nonmagnetic zinc or cadmium finish.

5.3 *Reproducibility of Optimum Magnetic Properties*

An important factor in determining the reproducibility of magnetic latching characteristics is the sensitivity of the hardened steel core to variations from the optimum physical hardness value, during manufacture. This effect is indicated by the latching curves of Fig. 8(a). These curves are representative of the data obtained with cores in the numbers 58 to 62 Rockwell (30-N) hardness range, and with magnet assemblies having the expected range of quality in parts and alignment. The latching curves show that the physical structures in the high-carbon steel cores, as obtained by heat treatments producing Rockwell hardness

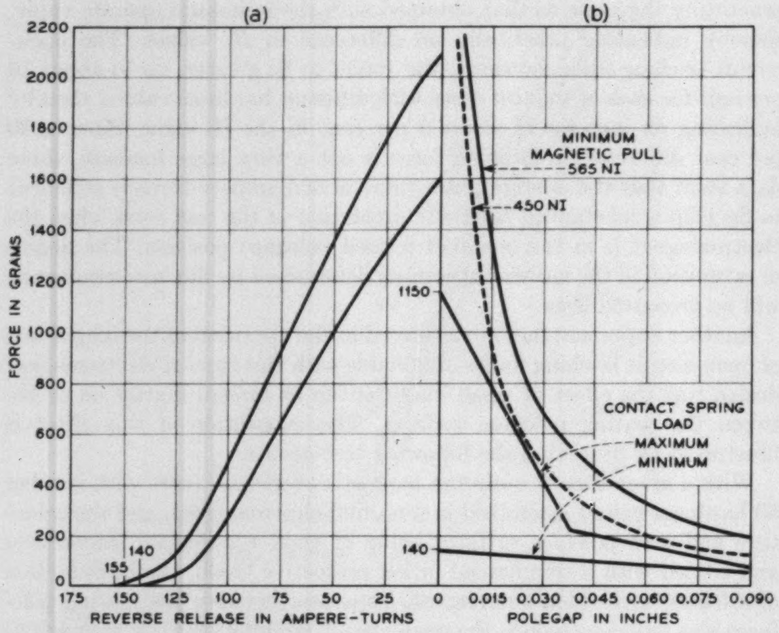


Fig. 8 — (a) Normal range of magnetic latching characteristics with 565 NI operating pulse and hardness of cores ranging from 58 to 62 (Rockwell 30-N); (b) extreme load and minimum operating pull characteristics.

readings of 58 to 62 on the 30-N scale, yield a combination of magnetic properties that provides satisfactory margins for the required magnetic latching function.

Tests were made also to determine the stability of the operating and latching properties from the standpoint of magnetic aging on the magnet cores. After about 200 hours of heating at a temperature of 100°C, no significant change due to aging could be detected.

Fig. 8(b) shows the operate pull curves for the same test parts and assemblies. The magnetic pull curve obtained with the minimum operating pulse strength of 565 NI shows that the open-gap tractive force is always considerably greater than the contact spring load, as the armature moves from the maximum-open polegap to the closed polegap position. The force differential between the load and the 565 NI pull, at each instantaneous value of polegap, determines the armature travel time. This time value, plus the time required for the current to build up to the just-operate value that starts the travel, is the maximum total operating time of the electromagnet. Since operating or switching times are of great importance, it is of interest to observe the following time data.

5.4 *Switching Times with Magnetic Latching Hold Magnet*

Fig. 9 represents a typical oscillogram of the operating-time characteristic of the magnetic latching hold magnet as it functions in a crossbar switch when load and operating power conditions are as follows: The contact spring load is the heaviest that may be encountered in the remote unit of a line concentrator; the circuit voltage is at the minimum value of 22 volts; and the circuit resistance is at the maximum value that provides the steady state value of 565 NI.

At zero time, two select magnets are energized simultaneously. At 0.030 second, the dip in the curve shows that the two associated select bars have rotated and interposed two wire fingers between the test hold magnet armature and crosspoint contacts on two separate horizontal levels. At 0.044 second, the test hold magnet is energized by the worst circuit current pulse. At 0.077 second, the test hold magnet armature has completed the switching of all contacts in the two crosspoints and in the HON spring assembly and has just reached the core poleface. At 0.122 second, the current has just reached about 95 per cent of its ultimate steady-state value.

These test values therefore show that the hold magnet operate time is a maximum of 0.033 second, and that the time required by the minimum circuit energizing pulse to build up the magnetic induction in the core

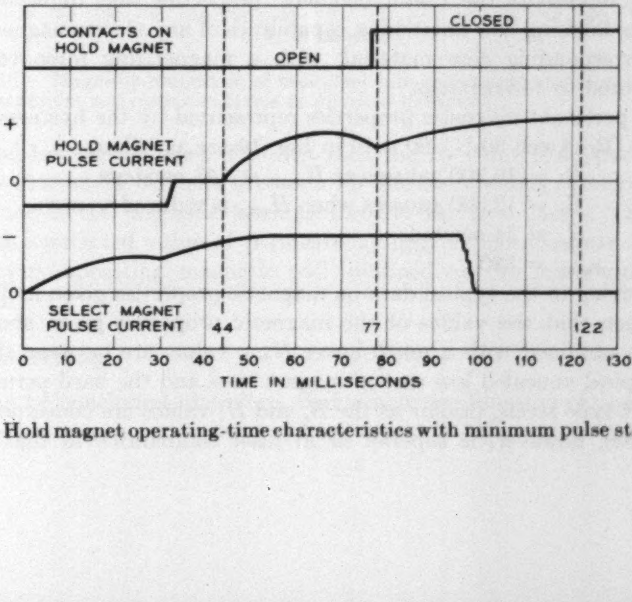


Fig. 9 — Hold magnet operating-time characteristics with minimum pulse strength.

to about 95 per cent of its steady-state value is 0.078 second. The latter time value represents a satisfactory margin, since 0.100 second was set up as the minimum time for the duration of the 22-volt energizing pulse. In this connection, it should be noted that the energizing pulse time can be reduced to much lower values by simply using higher operating voltage values to speed up the build up time of the energizing current to the same saturating value. For example, the pulse time required to obtain the same latching force capability was found to be only 0.015 second when the circuit voltage was increased to 90 volts and the ohmic resistance of the magnet circuit was increased to limit the steady-state current value to 0.200 ampere.

VI. FLUX MEASUREMENTS ON NEW MAGNET CORE

In order to observe the magnetic properties of the new magnet core material by itself, each of the core specimens representing the eight different levels of physical hardness was measured for its *B-H* magnetization characteristics. The measurements were made on a Bell Telephone Laboratories Cioffi recording flux meter system, which employs a Chattock magnetic potentiometer and an *H* integrator to measure and record the applied magnetizing force on the 3.5-inch-long test core specimen while it is being magnetized by the field between the poles of an electromagnet. The flux density *B* in the test specimen is measured and recorded directly from the test search coil on the specimen and the *B* integrator part of the system. Fig. 10 shows three of the *B-H* hysteresis loops so obtained. Each loop is designated by the hardness number of the steel core test specimen. The portion of each loop that lies in quadrants I and II represents the useful magnetic properties that determine the operate, latching and unlatching capabilities of any electromagnet using the corresponding core material, with a magnetizing force value of $H_{\max}$ equal to 143 oersteds.

The pertinent magnetic properties represented by the hysteresis loop for the (Rockwell 30-N) 60 steel in Fig. 10 are as follows:

B_s = 16,300 gausses at $H_{\max}$ = 143 oersteds,
B_r = 13,300 gausses when $H_{\max}$ is reduced to zero,
H_c = 24 oersteds,
$\mu_{\max}$ = 320.

Referring to the typical data on magnetic properties given in Table I, it is seen that the values of the magnetic properties given above, although obtained with a much lower $H_{\max}$ value, are between those of the typical annealed low-carbon magnet steel and the hard permanent-magnet-type steels, insofar as the B_s and H_c values are concerned. The B_r value, however, is superior or at least comparable to that of the

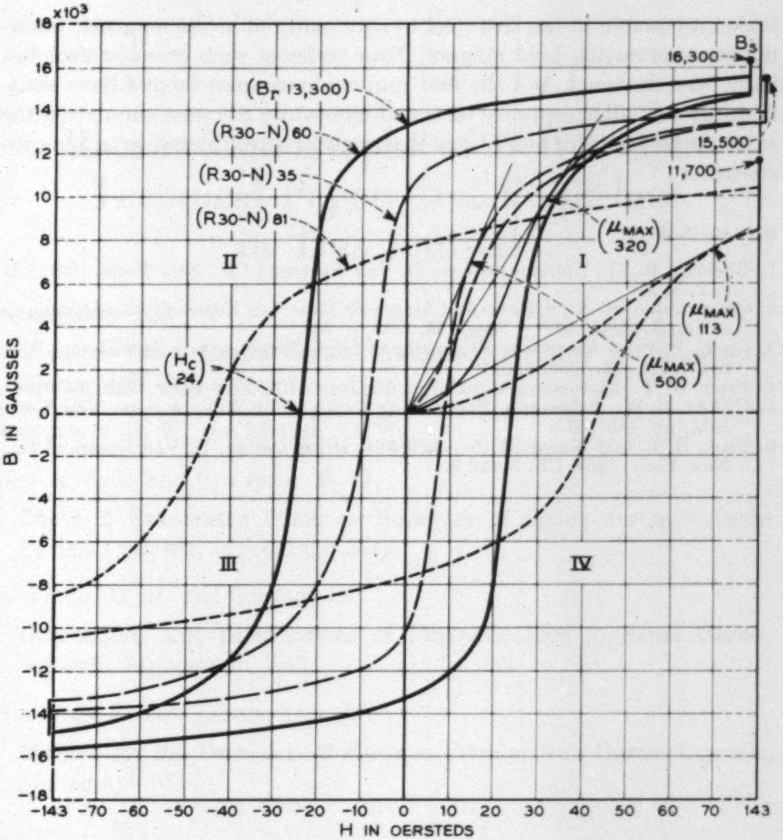


Fig. 10 — Magnetic properties of tool steel core specimens representing minimum, maximum and optimum levels of physical hardness.

annealed low-carbon steel. The fact that the B_r value is supported by an H_c value of 24 oersteds accounts for the outstanding strength and endurance of the magnetic latching force in the new design. The fact that the associated value of permeability $\mu_{\max}$ is 320 accounts for the satisfactory operating magnetic pull obtained on the minimum pulse strength.

VII. CONCLUSION

It can be concluded, therefore, that a new combination of coexisting values of magnetic properties has been found in high-carbon steel that makes it possible to use this steel as core material in the magnetic latching crossbar switch hold magnet. Four codes of such crossbar switches have been designed, and the first tool-made samples thereof have satisfactorily met all acceptance tests and laboratory life tests simulating the extreme field service conditions that may be encountered in a line concentrator.

BIBLIOGRAPHY

1. Bozorth, R. M., *Ferromagnetism*, D. Van Nostrand Co., New York, 1951, Ch. 1, 2, 3, 9.
2. Chegwidden, R. A., A Review of Magnetic Materials Especially for Communication Systems, Metal Prog., **54**, 1948, p. 705.
3. Cioffi, P. P., A Recording Fluxmeter of High Accuracy and Sensitivity, Rev. Sci. Instr., **21**, 1950, p. 624.
4. Frier, W. T., *Elementary Metallurgy*, McGraw-Hill, New York, 1942, pp. 66–89.
5. Gillett, H. W., *Behavior of Engineering Metals*, John Wiley & Sons, New York, 1951, pp. 135–159.
6. Peek, R. L. and Wagar, H. N., *Switching Relay Design*, D. Van Nostrand Co., New York, 1955, Ch. 3 and 9.

Recent Monographs of Bell System Technical Papers Not Published in This Journal*

ABRAHAMS, S. C.

Cryostat for a Single Crystal Automatic Neutron Diffractometer, Monograph 3546.

ANDERSON, E. W., see McCall, D. W.

BEACH, A. L. and GULDNER, W. G.

Effect of Evaporated Films on Recovery of Gases during Vacuum Fusion Analyses, Monograph 3547.

BOZORTH, R. M. and GELLER, S.

Interactions and Distributions of Magnetic Ions in Some Garnet Systems, Monograph 3548.

BUEHLER, E. and TANENBAUM, M.

Method for the Detection of Flaws in Yttrium Iron Garnet Crystals, Monograph 3549.

CIOFFI, P. P.

Relation between Permanent Magnet Configuration and Performance, Monograph 3550.

COLLINS, R. J. and KLEINMAN, D. A.

Infrared Reflectivity of Zinc Oxide, Monograph 3551.

COMPTON, K. G.

Sources of Underground Corrosion Potential Differences, Monograph 3552.

---

* Copies of these monographs may be obtained on request to the Publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. The numbers of the monographs should be given in all requests.

Danielson, W. E., see McDowell, H. L.

Danylchuk, I. and Katz, D.

Magnetic Amplifier Binary-to-Analog Conversion, Monograph 3553.

De Monte, R. W.

Synthesis of Cable Simulation Networks, Monograph 3487.

Douglass, D. C., see McCall, D. W.

Geller, S., see Bozorth, R. M.

Gibbons, D. F.

Thermal Expansion and Grüneisen Factor of Vitreous Silica, Monograph 3554.

Gross, F. J.

Simulation of Data-Switching Systems on a Digital Computer, Monograph 3556.

Guldner, W. G., see Beach, A. L.

Helfand, E. and Kirkwood, J. G.

Theory of the Heat of Transport of Electrolytic Solutions, Monograph 3555.

Hopfield, J. J., see Thomas, D. G.

Katz, D., see Danylchuk, I.

Kirkwood, J. G., see Helfand, E.

Kisliuk, P.

Calorimetric Heat of Adsorption—Nitrogen on Tungsten, Monograph 3557.

Kleinman, D. A., see Collins, R. J.

Laudise, R. A. and Sullivan, R. A.

Pilot Plant Production of Synthetic Quartz, Monograph 3558.

Logan, R. A. and Peters, A. J.

Impurity Effects Upon Mobility in Silicon, Monograph 3560.

McCall, D. W., Douglass, D. C. and Anderson, E. W.

Diffusion in Liquids, Monograph 3561.

McClusky, E. J., Jr.

A Comparison of Sequential and Iterative Circuits, Monograph 3562.

McDowell, H. L., Danielson, W. E. and Reed, E. D.

A Half-Watt CW Traveling-Wave Amplifier for the 5–6 Millimeter Band, Monograph 3563.

Moriya, T.

Theory of Magnetism of NiF_2, Monograph 3565.

Milner, P. C.

Interpretation of Measurements of Potential Decay on Open Circuit, Monograph 3564.

Peter, M.

Paramagnetic Resonance and Crystal Field in Two Nickel Chelate Crystals, Monograph 3566.

Peters, A. J., see Logan, R. A.

Porbansky, E. M., see Trumbore, F. A.

Priebe, H. F., Jr., see Shafer, W. L., Jr.

Radin, P. D., see Senitzky, B.

Reed, E. D., see McDowell, H. L.

Scheinman, A. H.

Numerical-Graphical Method in the Design of Multiterminal Switching Circuits, Monograph 3495.

Senitzky, B. and Radin, P. D.

Silicon p-n Junctions: Breakdown Characteristics; Electron Emission, Monograph 3567.

Shafer, W. L., Jr., Toy, W. N. and Priebe, H. F., Jr.

A Small High-Speed Transistor and Ferrite Core Memory System, Monograph 3568.

SHIMMIN, E. R., VANDERLIPPE, R. A. and WHITMAN, A. L.

A Small Automatic Teletypewriter Switching System, Monograph 3569.

SODEN, R. R., see Van Uitert, L. G.

SULLIVAN, R. A., see Laudise, R. A.

TANENBAUM, M., see Buehler, E.

TARTAGLIA, A. A., see Trumbore, F. A.

THOMAS, D. G. and HOPFIELD, J. J.

Exciton Spectrum of Cadmium Sulphide, Monograph 3570.

TOY, W. N., see Shafer, W. L., Jr.

TRUMBORE, F. A., PORBANSKY, E. M. and TARTAGLIA, A. A.

Solid Solubilities of Aluminum and Gallium in Germanium, Monograph 3571.

VANDERLIPPE, R. A., see Shimmin, E. R.

VAN UITERT, L. G. and SODEN, R. R.

Single-Crystal Tungstates for Resonance and Emission Studies, Monograph 3572.

WANNIER, G. H.

Wave Functions and Effective Hamiltonian for Bloch Electrons in an Electric Field, Monograph 3573.

WHITMAN, A. L., see Shimmin, E. R.

WINDELER, A. S.

Design of Polyethylene-Insulated Multipair Telephone Cable, Monograph 3574.

WOLFF, P. A.

Effects of Electron Correlation on the Optical Properties of Metals. Monograph 3575.

Contributors to this Issue

Václav E. Beneš, A.B., 1950, Harvard College; M.A., Ph.D., 1953, Princeton University; Bell Telephone Laboratories, 1953—. Mr. Beneš has been engaged in mathematical research on stochastic processes, traffic theory and servomechanisms. In 1959–60 he was visiting lecturer in mathematics at Dartmouth College. Member American Mathematical Society, Association for Symbolic Logic, Institute of Mathematical Statistics, Mind Association, Phi Beta Kappa.

James L. Flanagan, B.S., 1948, Mississippi State University; S.M., 1950, and Sc.D., 1955, Massachusetts Institute of Technology; faculty, Mississippi State University, 1950–52; Air Force Cambridge Research Center, 1954–57; Bell Telephone Laboratories, 1957—. He has specialized in work on speech communication over narrow bandwidths, including studies of acoustical, physiological and psychological phenomena related to speech and speech perception. Fellow Acoustical Society of America; member I.R.E., Kappa Mu Epsilon, Sigma Xi, Tau Beta Pi.

L. E. Franks, B.S., 1952, Oregon State College; M.S., 1953, and Ph.D., 1957, Stanford University; Bell Telephone Laboratories, 1958—. He is engaged in analytical studies of techniques for improving the performance of data transmission networks. Member Sigma Xi.

E. N. Gilbert, B.S., 1943, Queens College; Ph.D., 1948, Massachusetts Institute of Technology; M.I.T. Radiation Laboratory, 1944–46; Bell Telephone Laboratories, 1948—. Mr. Gilbert has been engaged in studies of information theory and switching theory. Member American Mathematical Society.

Bela Julesz, Dipl. in Electrical Engineering, 1950, Budapest (Hungary) Technical University; Kandidat in Technical Sciences, 1956, Hungarian Academy of Sciences; Bell Telephone Laboratories, 1956—. He was first engaged in studies of systems for reducing television bandwidth. At present, Dr. Julesz is working in visual research, particularly on problems of depth perception and pattern recognition. Member A.A.A.S., I.R.E.

C. Y. Lee, B.E.E., 1947, Cornell University; M.S.E.E., 1949, and Ph.D., 1954, University of Washington; instructor in electrical engineering, University of Washington, 1948–51; Bell Telephone Laboratories, 1952—. Mr. Lee has been engaged in studies of mathematical problems arising from computers and digital systems. He was a visiting member of the Institute for Advanced Study in the School of Mathematics during the academic year 1957–58. Member American Mathematical Society, I.R.E., Eta Kappa Nu, Sigma Xi.

Irwin W. Sandberg, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1958—. He has been concerned with analysis of military systems, particularly radar systems, and with synthethis and analysis of active and time-varying networks and linear array antennas. Member I.R.E., Eta Kappa Nu, Sigma Xi, Tau Beta Pi.

David Slepian, University of Michigan, 1941–43; M.A., 1947, and Ph.D., 1949, Harvard University; Bell Telephone Laboratories, 1950—. He has been engaged in mathematical research in communication theory, switching theory and theory of noise, and has been mathematical consultant on various Bell Laboratories projects. In 1958 and 1959 he was Visiting Mackay Professor of Electrical Engineering at the University of California at Berkeley. Member A.A.A.S., American Mathematical Society, Institute of Mathematical Statistics, I.R.E., Society for Industrial and Applied Mathematics, U.R.S.I. Commission 6.

A. W. Warner, B.A., 1940, University of Delaware; M.S., 1942, University of Maryland; instructor in physics, Lehigh University, 1941-42; Western Electric Co., 1942–43; Bell Telephone Laboratories, 1943—. Since joining the Bell System Mr. Warner has been continuously engaged in the development of high-frequency quartz crystal units. At present he is continuing the development of very stable crystal units and other solid state devices making use of crystalline quartz. Senior member I.R.E.

Frank A. Zupa, B.S. in E.E., 1922, Cooper Union; Western Electric Co. Engineering Dept., 1918–25; Bell Telephone Laboratories, 1925—. He was engaged in evaluation testing of materials and switching apparatus for about six years, and in design and development engineering work on practically all types of telephone relays for more than 30 years. During World War II he was in charge of the packaging design for production of the optical proximity fuse and in the evaluation testing of magnetic mine M11. At present he is in charge of a group engaged in new switch and relay design developments.